THE UNIVERSITY
of EDINBURGH

Germline variants associated with alternative splicing in colonic mucosa

THE UNIVERSITY of EDINBURGH

Toby Gurran

Thesis submitted for the degree of

Doctor of Philosophy

School of Medicine and Veterinary Medicine

The University of Edinburgh

2019

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

MRC Human Genetics Unit

CANCER RESEARCH UK

Declaration

I declare that this thesis was composed entirely by myself and that the research presented is my own unless otherwise stated. No part of this research has been submitted for any other degree or professional qualification.

Toby Gurran

September 2019

Abstract

The heritability of colorectal cancer (CRC) has been estimated between 7.4% and 26% from a range of analyses based on family lineages and genetic similarity. Certain rare, high penetrance variants are well characterized, though these are estimated to account for only ~5% of all CRC cases. The majority of GWAS-identified risk SNPs for CRC fall within non-coding regions, and the mechanisms by which the majority of these variants contribute to disease predisposition are yet to be elucidated. However, recent studies have highlighted the contribution of alternative splicing to cancer progression, and have linked variants altering splicing patterns to predisposition to other complex traits.

This study has analysed RNA-seq from 221 samples of colonic mucosa (the precise tissue of origin of CRC) from a Scottish cohort to identify variants associated with quantitative changes in the splicing patterns of genes (sQTLs). All individuals were genotyped from blood samples via SNP-chips, and imputation increased the number of testable variants to 4 million. Transcript expression was quantified with the alignment-free Salmon algorithm. Two separate approaches with complementary methodologies were used to identify sQTLs: the sQTLseekeR package which analyses whole transcripts, and the Leafcutter package which infers changes in intron usage. Between the two, over 15,000 variants were identified as corresponding to changes in the ratio of expression of transcripts or the ratio of intron excision from over 6,800 protein-coding and lncRNA genes. Effect size and expression thresholds were applied to retain only the top 8% most likely functionally relevant sQTLs.

The thresholded sQTLs were found to be enriched in peaks of active chromatin marks, DNase accessible regions and putative regulatory elements, relative to a population of 100,000 non-sQTL SNPs sampled from the same search windows and with the same proportions of minor allele frequencies as the sQTL SNPs. They were similarly enriched within regions predicted to be active from probabilistic deconvolution of signals from multiple histone marks constructed by the Roadmap Epigenetics Consortium. sQTLs were enriched within linkage blocks containing eQTLs (expression quantitative trait loci) identified from the same cohort, and eQTLs identified from GTEx sigmoid and transverse colon tissues; however the lead SNPs associated with sQTLs and eQTLs were different in 97% of cases, implying a

strong degree of independence between the two classes of event. Thresholded sQTL variants identified by the Leafcutter package were found to be significantly enriched within a meta-GWAS for CRC consisting of 20,818 cases and 37,822 controls. Between both packages, sQTLs were found for 9 genes associated with CRC in the NHGRI-EBI GWAS catalog, 4 genes curated in the COSMIC database as relevant to CRC progression, and a further 29 oncogenes or tumour suppressors implicated in any cancer.

Together these observations imply that the alteration of patterns of transcript expression in the colonic mucosa mediated by germline SNPs is one of the genetic mechanisms underpinning predisposition to CRC. The sQTLs identified herein could be further used in colocalisation analyses to fine-map GWAS causal variants, and in transcriptome wide association studies (TWAS) to identify new CRC predisposition loci.

Lay summary

DNA stores instructions within cells, telling them when to grow and how quickly to do so. The majority of cancers are caused by mutations in a person's DNA which lead cells to multiply and spread too quickly. Most mutations which drive cancer are accumulated incidentally throughout a person's lifetime, either through mistakes made in copying DNA whilst making new cells, or from exposure to mutagens such as UV light or cigarette smoke.

However, some people are born with certain mutations already in their DNA which make them more predisposed to develop a particular cancer over the course of their lifetime. There are certain mutations like this which are well known to greatly increase the likelihood of developing cancer of the colon, and people with these conditions can first develop tumours from a very early age and have a poor life expectancy. These such mutations are very rare in the general population however, and there are other more common mutations which less strongly predispose to colorectal cancer - though when combined together can still make a significant contribution to an individual's likelihood of developing it.

Genes are individual sequences of DNA which send a specific instruction to the cell. This means that when a mutation occurs within the sequence of a gene, it is relatively easy to predict the consequence it will have by translating how the instruction has been changed. The rare, high-impact mutations predisposing to colon cancer tend to occur within genes, however the majority of the more common and lower-impact mutations occur in regions outside of genes - so their consequences are less easy to predict.

This project aimed to investigate a certain class of these DNA variants termed "alternative splicing quantitative trait loci", which do not change the way genes are coded, but change the degree to which different sub-sections of genes are expressed. This required decoding the DNA of many individuals in combination with analysing their RNA from the specific tissue where colon cancer develops; the colonic mucosa. RNA is the way that instructions encoded in genes are delivered from DNA storage out of the nucleus into the cell body. The hypothesis of this thesis is that "alternative splicing quantitative trait loci", which alter the amounts by which different portions of the RNA messages are sent out, may account for some of the common, low-impact DNA variations which can predispose people to colon cancer.

Acknowledgements

Thank you to CRUK for funding this PhD, and to the University of Edinburgh and the MRC for funding the IGMM. It has been a pleasure to work at an institute with such a collaborative, friendly and diverse community of internationally recognised scientists.

I am very grateful to my supervisors Dr Colin Semple, Dr Susan Farrington, and Dr Alison Meynert. Thank you to Colin for giving me the opportunity to undertake this PhD, and for advice and guidance throughout this project. Thank you to Susan for her support and for welcoming me into the weekly lab meetings. I am indebted to Alison for her generous and compassionate supervision, for always having the time to share her experience with me, and for her tireless reviewing of thesis drafts.

I am grateful to the surgical teams who collected the samples that provided the data for this project, in particular Dr Li Yin Ooi and Dr Peter Vaughan-Shaw, under the supervision of Professor Malcolm Dunlop. I wish to sincerely thank all the individuals who selflessly donated samples.

Sincere thanks go to Dr Vicky Svinti and Dr Maria Timofeeva for facilitating my access to the genotyping and expression data, and for their generous help in assisting and advising me with analyses. Thanks also to Graeme Grimes for his coding wisdom.

Thank you to all members of the Evogen Group, the Biomedical Genomics Section, and everyone at the institute who has attended one of my presentations or given me comments or feedback.

I am very appreciative of the IGMM and University of Edinburgh IT Teams for maintaining the “Eddie” compute cluster and for all their work behind the scenes.

I wish to thank all of the many great friends I have made whilst in Edinburgh. I have wonderful memories sharing laughs at dinner parties, restaurants, the festival, and of course the inimitable hospital canteen. It has been a pleasure taking this journey with you, and I look forward to many happy reunions in the future.

Finally, I wish to give the deepest thanks to my incredible parents and family for all their love and support throughout my studies. I am eternally grateful to have you.

List of Abbreviations

- AJCC: The American Joint Committee on Cancer
- ALL: Acute Lymphocytic Leukaemia
- AML: Acute Myeloid Leukaemia
- Bigwig: A file format representing the density of reads aligned to a particular sequence which saves memory by not requiring the visualisation of each individual read.
- BMI: Body Mass Index
- B-NHL: B-cell Non-Hodgkin Lymphoma
- BPS: Branchpoint Sequence (spliceosome complex binding site)
- CAR-T: Chimeric Antigen Receptor T cell
- CCG Group: Colorectal Cancer Genetics Group. Team of surgeons and scientists based at the Western General Hospital and IGMM Edinburgh led by Prof Malcolm Dunlop and Dr Susan Farrington.
- CDS: Coding Sequence, the region of DNA that is translated to form proteins
- CEU
- ChIP-seq: Chromatin Immuno-Precipitation followed by sequencing
- CI: Confidence Interval
- CLL: Chronic lymphocytic leukaemia
- CRC: Colorectal cancer
- DGN: Depression Genes and Networks Cohort
- DLPFC: Dorsolateral Prefrontal Cortex brain tissue
- DM: Dirichlet Multinomial distribution
- ENCODE: The Encyclopedia of DNA Elements
- ESE: Exonic Splicing Enhancer
- ESS: Exonic Splicing Suppressor
- ETP-ALL: Early T-cell Precursor Acute Lymphoblastic Leukaemia
- FDR: False Discovery Rate
- FLD: Fragment length distribution
- FPKM: Fragments per kilobase of sequence per million mapped reads
- GATK: Genome Analysis Tool Kit from the BROAD Institute
- GBM: Glioblastoma Multiforme
- GEAUVADIS populations: CEU (CEPH) are Utah with European ancestry, Finns (FIN), British (GBR),Toscani (TSI) and Yoruba (YRI)
- GTEx: Genotype Tissue Expression Project
- GWAS: Genome Wide Association Study
- HNSCC: Head and Neck Squamous Cell Carcinoma
- ISE: Intronic Splice Enhancer
- ISS: Intronic Splice Suppressor
- INDEL: Insertion or Deletion
- KIRC: Kidney renal clear cell carcinoma
- LCLs: Lymphoblastoid Cell Lines. Immortalized (usually be means of inoculation with Epstein-Barr virus) cell lines derived from human white blood cells.
- Limma: Linear modelling of microarrays
- M: Million
- MARD: Mean Absolute Relative Difference
- MBL: Monoclonal B-lymphocytosis, a precursor disorder to CLL
- MD: “Maximum Difference”. A value assigned by the sQTLseekeR algotirhm to indicate the effect size of an sQTL event, corresponding to the difference in

relative expression between the two transcripts with the greatest reciprocal change of relative expression between homozygous reference and homozygous alternative genotype groups.

- MHC: Major Histocompatibility Complex region of Chromosome 6.
- MR: Medelian Randomization.
- mRNA: Messenger RNA
- NHS: National Health Service
- NMD: Nonsense Mediated Decay pathway
- NSCLC: Non Small Cell Lung Cancer
- ONT: Oxford Nanopore Technologies
- OR: Odds Ratio (e.g. as in a Fisher's Exact Test)
- PBMC: Peripheral blood mononuclear cell
- PEER: Probabilistic Estimation of Expression Residuals
- PRS: Polygenic Risk Score
- PSI: Percent Spliced In. A measure of alternative splicing based on the ratio of presence or absence of an intron or an exon of a transcript.
- PTC: Premature Termination Codon
- QC: Quality Control
- Q-PCR: Quantitative Polymerase Chain Reaction
- RBP: RNA Binding Protein
- RNA: Ribonucleic Acid
- RNP: Ribonuclear Protein
- RR: Relative Risk
- S: Svedberg. A unit denoting the rate of sedimentation of a particle when centrifuged.
- SAVs: Splicing Associated Variants
- SCC: Squamous Cell Carcinoma
- SCOVIDS: Scottish Vitamin D Study
- SE: Standard Error
- SMR-HEIDI: Summary data-based Mendelian randomization with heterogeneity in dependent instruments.
- SNP: Single Nucleotide Polymorphism.
- snRNA: Small nuclear RNA
- SNV: Single Nucleotide Variant.
- SOCCS: Scottish Colorectal Cancer Susceptibility
- sQTL: Alternative splicing Quantitative Trait Locus
- SR: Serine/Arginine-rich proteins. Bind to ESEs to promote splicing.
- SS: Splice Site
- SV: Structural Variant
- SVA: Surrogate Variable Analysis
- svQTL: Splicing variance Quantitative Trait Locus
- TA: Transit amplifying cell
- T-ALL: T-cell Acute Lymphoblastic Leukaemia
- TNM: AJCC system for scoring cancer growth based on tumour size, number of affected lymph nodes and presence or absence of distal metastases.
- TOM: Topological Overlap Matrix
- TSG: Tumour Suppressor Gene
- TSS: Transcription Start Site
- UTR: Untranscribed Region. Can be 5' or 3' relating to the transcript in question.
- VEP: Variant Effect Predictor

- WES: Whole Exome Sequencing
- WGCNA: Weighted gene correlation network analysis
- WGS: Whole Genome Sequencing
- WT: Wild Type
- WTCRF: Wellcome Trust Clinical Research Facility

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Colorectal cancer and its causes

Colorectal cancer (CRC) is the fourth most common cancer in the UK[1], and is the second leading cause of cancer-related mortality[2]. The disease can be classified into four stages of increasing severity, and the majority of cases of CRC are only diagnosed at a late stage in the United Kingdom[3]. The American Joint Committee on Cancer (AJCC) classifies CRC into four different stages based on the size of the tumour, the number of affected lymph nodes and number of distal metastases, termed the "TNM" system[4]. Survival is high for patients diagnosed early in the stages of disease progression with over 94% five-year-survival for stage 1 diagnoses, as compared to less than 9% five-year-survival for patients diagnosed at stage 4[5,6]. Screening programmes to detect faecal occult blood as a marker of CRC presence reduce relative risk of CRC by up to 25%(RR 0.75, 95% confidence interval 0.66-0.84)[7]; however only 10% of cases are currently diagnosed in this way[8]. Understanding the causes of CRC predisposition is therefore particularly important as this would enable patient stratification to target screening and early detection resources to individuals at the highest risk[9].

1.1.1 Physiology of the colon

The large intestine is separated into two sides according to position relative to the splenic flexure (*Figure 1.1*). The right side contains the caecum, appendix, ascending colon and transverse colon. The left side constitutes the descending colon and sigmoid colon, which after the rectosigmoid junction marks the transition into the rectum and anus. Together the large intestine and rectum comprise the colorectum. The two sides of the colorectum originate from different subcellular populations during embryogenesis: regions on the right of the splenic flexure originate from the midgut and those on the left from the hindgut[10].

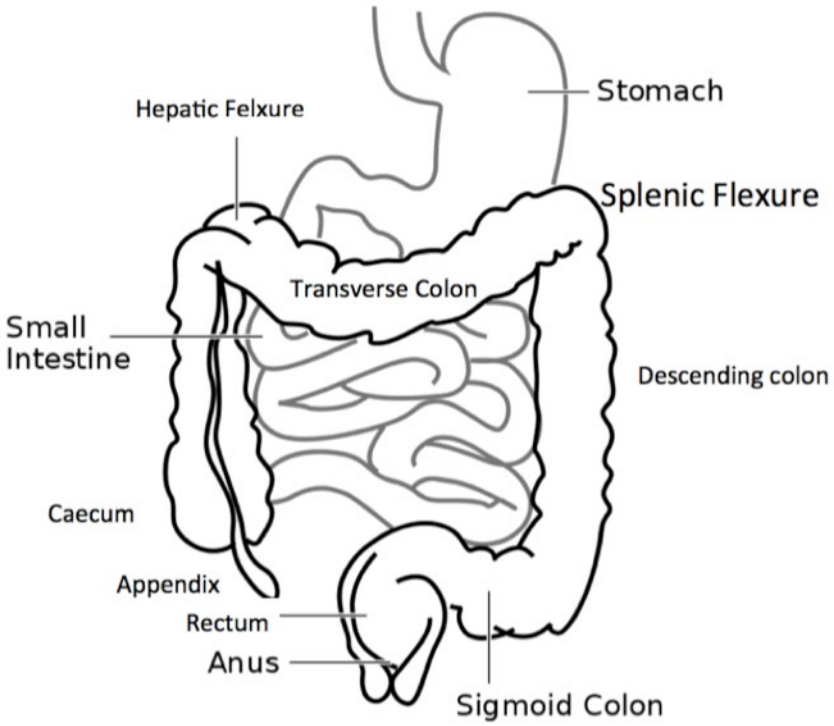


Figure 1.1 Colon physiology. Adapted from Wikimedia Commons[11].

The gastrointestinal tract is separated into 4 distinct layers: mucosa, submucosa, muscularis externa and serosa (*Figure 1.2*). The colonic mucosa is further separated into 3 key layers: columnar surface epithelial cells, the lamina propria (connective tissue) and an outer muscular layer (muscularis mucosa)[10]. Simple tubular glands are formed by invaginations of the epithelial cells into the mucosa and are termed “crypts of Lieberkühn”. The crypts increase surface area for the absorption of water and solutes in the large intestine, and also constitute the structure for epithelial cell turnover.

The base of each crypt is populated by intestinal stem cells[12] which are able to differentiate into any of the cell types specific to the colonic epithelium, including: absorptive enterocytes, goblet cells which secrete mucins to coat the intestine and protect against pathogens[13], enteroendocrine cells which act as chemoreceptors and secrete gastrointestinal hormones in response to the luminal environment[14], paneth cells which secrete antimicrobial defensin proteins and lysosomes[15] and tuft cells which produce intestinal opioids[16] and aid in activation of the type-2 cytokine immune response of paneth cells against intestinal parasites such as helminth worms[17].

The submucosa is a layer of connective tissue containing nerves (Meissner's plexus) and blood vessels[10]. The muscularis layer contains an inner circular and outer longitudinal sheet of muscle which allow for the motility of the coloretum via peristalsis. The outermost layer of the colorectum constitutes the peritoneum which contains blood vessels for the support of the abdominal organs and binds them to each other and the walls of the abdominal cavity[18].

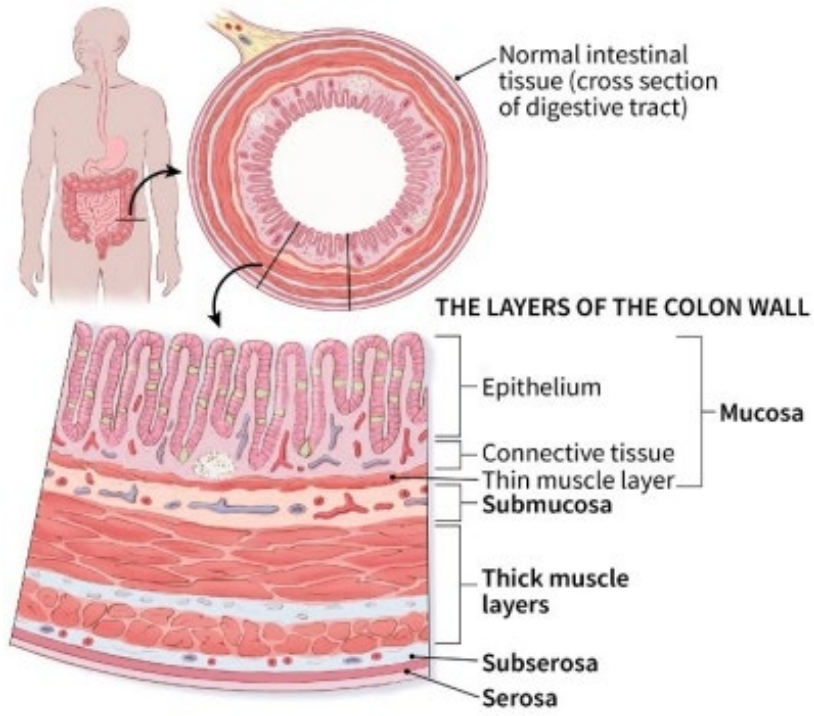


Figure 1.2 Layers of colon wall. From American Cancer Society (ACS)[19].

1.1.2 The stem cell theory of cancer

Intestinal epithelial cells at the surface of crypts are rapidly turned over and sloughed off on average every 5-7 days[20]. This requires rapid and continuous re-population of differentiated epithelial cells from $LGR5^+$ stem cells at the base of crypts, which are dependent on Wnt ligands for maintenance of their stem-cell state[21] (*Figure 1.3*). Whilst the requirement of paracrine Wnt has been known for over a decade[22], the exact source had remained unknown until it was discovered in 2018 that GLI1-expressing subepithelial mesenchymal cells were necessary for

intestinal epithelial renewal in mice[23]. The stem cells are also fed BMP agonists including GREM1, GREM2 and CHRDL1 from subepithelial myofibroblasts and smooth muscle cells below, which serve to maintain the required signalling axis for crypts to form in the appropriate orientation[24]. The signalling gradient changes as cells progress up the axis with terminal differentiation being induced by activation of the TGF beta pathway via removal of restrictions on BMP ligands[25], along with activation of components of the Notch pathway[26], Myc signalling network[27,28] and increased release of ephrin ligands[29,30].

It is theorised that the stem cells are the population from which colorectal cancer initiates as they are the only cells that have the potential to produce all other epithelial cell types and which persist long enough to acquire the genetic mutations or epigenetic aberrations necessary to fulfil the multi-hit hypothesis[31]. The cancer stem cell hypothesis purports that there is then a latent minority population of stem cells which retain plasticity and serve to re-establish tumours following development of resistance to chemotherapy, radiotherapy or targeted agents[32–34]. $LGR5^{+}$ stem cells were proven to be the cell of origin for intestinal tumours when APC was selectively deleted in them in a mouse model, which generated macroscopic adenomas within 3-5 weeks[35]. This was in contrast with shorter-lived transit amplifying cells which when subjected to the same mutation failed to develop tumours other than microadenomas which stalled in growth shortly after induction[35]. There exists a second pool of rarer intestinal stem cells which express BMI1 and inhabit a region encompassing the 4 cells distal to $LGR5^{+}$ cells at the bases of crypts[36,37], and they have been shown to be able to regenerate $LGR5^{+}$ cells upon artificial ablation of crypt bases as a simulation of intestinal damage[38]. These cells are also able to initiate tumours if exposed to the necessary genetic/epigenetic perturbations[36].

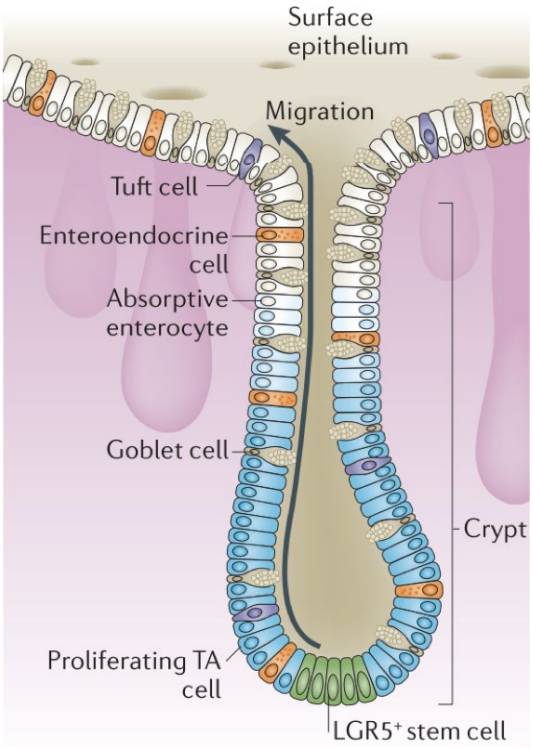


Figure 1.3 Stem cell renewal of crypt structure. From Barker 2014[20]. TA: transit amplifying.

There is an opposing theory of "top-down" carcinogensis based on histological observations of early polyps appearing at the luminal surface of crypts with no contact to the stem cell compartment[39] which hypothesises that terminally differentiated cells may re-gain pluripotency through specific genetic or epigenetic lesions[40]. Constitutive NF-kB expression has been shown to upregulate Wnt signalling to a sufficient level to allow conversion of $LGR5^-$ cells to $LGR5^+$ cells with stem-like properties capable of tumour initiation upon concomitant mutation of *APC* and *KRAS*[41]. Whilst the two theories are not necessarily mutually exclusive, crypt-fission driven by the hyperproliferative progeny of mutated $LGR5^+$ stem cells is accepted as the primary mechanism by which early adenomas spread to neighbouring crypts[12,42].

1.1.3 CRC aetiology is influenced by physiology and gender

There are differences in epidemiology and etiology observed between CRC occurring in the two sides of the colon divided by the splenic flexure[43,44]. A study of 9,550 cases of colorectal cancer from Florida's statewide registry found right-sided

cancers had a higher incidence rate in females than males (55% vs 45%) and were on average identified at a later age than left-sided tumours (median 73.7 vs 69.4 years)[45]. A retrospective survival analysis of 77,978 patients found right-sided cancers to have poorer prognosis with a median survival of 78 vs 89 months, and a 5% greater hazard ratio for risk of mortality compared to left-sided cancers (1.04: 95% confidence interval, 1.02-1.07)[46]. In her thesis entitled "Post-GWAS functional characterisation of colorectal cancer risk loci", Dr Li Yin Ooi observed 55 genes differentially expressed >1.5 log-fold-change (12 > 2.0 log fold change) between right (39) and left (79) sided normal colonic mucosa samples via microarray[47]. The difference in survival is complicated by stage at diagnosis: right-sided tumours are less likely to be diagnosed early. A study of 3,552 Japanese colorectal cancer cases found 22% of left-sided cancers were diagnosed at stage 1 compared to just 15% of right-sided[48].

The fact that females have a higher incidence of right-sided CRC, which is less readily detected and has a poorer prognosis, may explain the poorer survival rates for females over the age of 65 compared to males[49]. However males of all ages are more likely to develop CRC than females. In the UK, 1 in 14 men will develop CRC over the course of their lifetime compared to 1 in 19 women[2].

1.1.4 Molecular pathology of CRC

Adenocarcinomas are cancers derived from epithelial cells, whilst sarcomas are derived from mesenchymal cells. 95% of colorectal cancers are adenocarcinomas, with the remaining percentage consisting of carcinoid tumours, squamous cell carcinoma, lymphoma and the rarest being sarcoma[50]. The AJCC TNM system classifies Stage 0 (or "*in situ*") CRC as tumours which are only growing on the most superficial luminal mucosa layer of the colon[4]. Stage I tumours have grown through the mucosa and the muscularis mucosa into the submucosa and potentially also into the muscularis propria but no further. Stage II tumours may have grown into the outermost layer of the colon and potentially through the colon wall into other organs but without spreading from the primary site. Stage III is defined by the spread of the cancer to lymph nodes and Stage IV is reached once any metastases in allosteric organs or tissues are observed.

Sporadic CRC can be divided into two main groups. 84% of the cases are classified as presenting “chromosomal instability” (CIN), mainly attributable to somatic copy number alterations and chromosomal translocations[51] which can be caused by telomere instability and/or improper chromosomal segregation during mitosis[52]. Though whilst termed unstable relative to other CRC cases, as a whole CRC is one of the least genomically unstabile cancers when compared with others such as high-grade serous ovarian cancer (HGCOC) and breast invasive carcinoma (BRCA)[53,54]. The other 16% of CRC cases present a hypermutated genome caused by microsatellite instability (MSI), which can be further subclassified into those caused by defects in DNA mismatch repair pathways (MMR: 13%) or those with DNA-polymerase Epsilon mutations (POLE: 3%)[55].

CIN tumours most often arise with an initial dysregulation of Wnt signalling. >80% of adenomas present loss of APC function[56] and a further ~10% have mutational or epigenetic inactivation of other components of the same pathway e.g. β-catenin[57]. Driver mutations, including in *KRAS* and *PIK3CA*, are then accumulated as the tumour progresses in size and dysplasia, followed by further loss of tumour suppressors including *SMAD4* and *TP53* which usually precede invasive growth beyond the wall of the colon and metastasis[55]. Almost 50% of CIN tumours possess constitutively activating *KRAS* mutations, in contrast to the *BRAF* mutations most commonly being responsible for driving the EGFR proliferative pathway in MSI tumours[58].

Sporadic MSI tumours most commonly initiate by promoter hypermethylation of *MLH1*, possibly by overexpression of DNMT3B[59], which leads to defective surveillance of single-base mismatches during DNA replication and a particular propensity for mutations in microsatellites of repeating units of 1-4bp[55]. Disruption of other genes which can lead to the phenotype include *MSH2*, *MSH6*, *PMS1* and *PMS6*[60], but interestingly not all CRC with the MSI phenotype have identified mutations in known MMR genes[61], which implies there may be uncharacterized facets to the DNA repair process[62]. A subset of MSI CRC tumours have ultramutator phenotypes due to mutations in the intrinsic exonuclease domains of DNA Polymerase Epsilon or Delta 1, *POLE* or *POLD1*, which are responsible for proofreading of base-substitution simultaneous with DNA replication[63]. The most common driver mutations for MSI CRC are constitutively activating mutations in *BRAF*[64] such as V600E which renders the catalytic cleft of BRAF permanently

accessible and can increase its activity by 500-fold, removing the cell's dependence on extracellular signals for activation of the MAPK proliferative pathway[65].

1.1.5 Consensus Molecular Subtypes of CRC

CIN tumours can be further subdivided based on their expression profiles, which are known to capture latent biologically relevant information about tumour characteristics[66]. The Colorectal Cancer Subtyping Consortium (CRCSC) used microarray profiling of 3,962 CRC samples from 18 datasets to perform supervised clustering into consensus subtypes of biological and clinical relevance[67]. Six groups had previously created subclassifications of CRC, with between three and six classes defined by each group. The consortium began by constructing a network whereby nodes were each of the 27 different sets previously created by these groups, and edges were the Jaccard similarity between set membership of each of the nodes. They applied a Markov Clustering algorithm to this network which resulted in four consensus molecular subtypes, named CMS1-4[68]. 3,104 (78%) of the samples were classified directly into these clusters by the Markov Clustering, so the authors used these as a gold-standard set to train a random forest algorithm to assign samples to subtypes based on gene expression. Applying this algorithm to the 858 originally unassigned samples they were able to unambiguously allocate a further 339 samples to a single subtype, with the remaining 519 samples (13% of total original samples) being unclassified. Those that were unclassified did not represent a further outgroup, but were intermediate or ambiguous between the 4 major subtypes, possibly reflecting intratumour heterogeneity or lower quality sequencing.

The majority of the cancers classified as MSI by the two-class system fell into subtype CMS1 which is characterized by a hypermutator phenotype and hypermethylation profiles, with a high frequency of *BRAF* constitutive activating mutations co-occuring with these characteristics[69]. Pathway analysis of genes most highly expressed in CMS1 tumours also revealed signatures of immune cell infiltration (specifically class 1 T helper cells and cytotoxic T cells), along with upregulation of genes involved in immune surveillance evasion[70].

The CIN tumours were deconvoluted into the remaining three classes. CMS2 had the most frequent instances of copy number gains and losses. CMS3 had fewer copy number variants, and more CpG island methylation than CMS2 or CMS4, and had the highest frequency of *KRAS* driver mutations. *APC* loss was significantly more frequent in CMS2-4 than CMS1. >25% of tumours across all groups had *TP53* mutations with the greatest amount (>75%) in CMS2. The fact that there were no clearly defined co-occurrences of key driver mutations in any of the CMS2-4 tumour subtypes indicates that simple genomic aberrations cannot be relied upon for dependable stratification of intrinsic tumour biology, and serves to highlight the power of transcriptional analyses and the latent features they can uncover. The molecular subtypes are clinically relevant: CMS1 tumours have a better prognosis than CMS2-4 tumours[71] and greater likelihood of showing clinical response to immunotherapies[72,73], likely due to the large number of non-self tumour-neoantigens that can be generated as a result of hypermutator phenotypes[74,75]. In contrast, there is a clinical need for multi-drug regimens for the CMS2-4 subtypes where there is rarely a single dominant driver gene[76].

1.1.6 Mendelian traits predisposing to CRC

Whilst the majority of CRC cases are sporadic, there are high-penetrance germline mutations which strongly predispose individuals to develop colorectal cancer, and other lower impact variants which together are estimated to account for approximately 20% of CRC cases[77]. The most potent inherited syndrome is Familial Adenomatous Polyposis (FAP), which if left unchecked leads to individuals developing hundreds to thousands of precancerous polyps throughout their large intestine. There are also extracolonic phenotypes associated with FAP which can include tumours in the small intestine, stomach, adrenal gland adenomas thyroid gland carcinomas, desmoid tumours arising from fibroblasts throughout the body and hypertrophy of retinal pigment epithelium in the eye[78].The incidence has been estimated to be between 1/8,000 to 1/14,000 live births[79,80], and sufferers have a 50% chance of developing adenoma by age 15 and 95% chance by age 35[81] with equal penetrance across genders. FAP is caused by mutations in the *APC* gene[82], a tumour suppressor which in the absence of Wnt signalling usually functions to facilitate the phosphorylation of beta-catenin by GSK3B, which results in it being tagged with ubiquitin for proteasomal degradation[83]. With APC inactivated, beta-catenin is constitutively free to transition into the nucleus, where as a coactivator of

T cell transcription factor (TCF) and lymphoid enhancer factor (LEF) it mediates expression of cell proliferative and pro-survival genes[84].

FAP is an autosomal dominant condition, however inheritance of two faulty copies of APC is embryonic lethal; in accordance with Knudson's two-hit hypothesis that individuals who inherit a single mutated APC allele then go on to lose the other in somatic cells[85]. The *APC* gene has a domain key to beta-catenin interaction, and whereas the majority of FAP mutations are truncating or nonsense, there is a hotspot "mutation cluster region" observed around codon 1300, likely as a result of selection pressure to retain the N-terminal function of APC whilst deactivating the C-terminus responsible for the beta-catenin degradation[86].

Lynch Syndrome is caused by mutations in the DNA mismatch repair pathway, and predisposes to colorectal cancer, with a 70% bias towards incidence of right-sided tumours[87]. Unlike sporadic CRC, in which the MSI/CMS1 phenotype is most commonly caused by promoter methylation of *MLH1*, Lynch syndrome is more evenly attributable between approximately 45% *MLH1* and 45% *MSH2* inactivating mutations[88]. The remaining 10% is made up predominantly of mutations to *MSH6*, *PMS1* or *PMS2*. The penetrance of germline mismatch repair genes is not as complete as in FAP, and also exhibits a gender bias, with estimates of between 74-94% penetrance in males and 30-63% in females[89,90]. Through faulty mismatch-repair Lynch syndrome also predisposes to endometrial cancer and carcinoma of the small intestine, stomach, ovary, kidney and breast[91]. Distinct to Lynch syndrome, MUTYH-associated polyposis is also caused by mutations in DNA-repair, specifically in the *MUTYH* gene involved in DNA base excision repair. It has a milder phenotype than *APC* mutations, and carriers of this recessive allele are classified as either having MUTYH-associated polyposis or "attenuated" FAP[92], which is a term that can also be applied to less severe cases of FAP with less damaging mutations which only lead to between 10-99 colonic adenoma polyps.

There are other rarer Mendelian diseases with incidences of between 1/30,000 to 1/200,000 in the population which predispose to CRC to a lesser extent than Lynch Syndrome or FAP. Peutz-Jeghers syndrome and Cowden syndrome are mediated by disruptive mutations to the tumour suppressor genes *STK11/LKB1* and *PTEN* respectively, both of which play roles in moderating the AKT-PI3K signal transduction pathway[93,94]. Juvenile polyposis is caused by mutations in either

SMAD4 or *BMPR1A*, both of which contribute to the TGF-beta signalling pathway, and SMAD4 also plays a role in tempering Wnt signalling[95].

1.1.7 Common risk variants predispose to CRC

Outside of these inherited syndromes, CRC is a complex trait which has both environmental and genetic components. GWAS studies have identified 79 loci significantly associated with increased risk of developing CRC in European cohorts[96], though each individual locus is relatively common and typically confers a low risk. The *TGFβR1*6Ala* mutation may be present in up to 14% of the Caucasian population and carries an odds ratio of 1.20 for CRC risk[97]. The SNP rs5934683 upstream of the *SHROOM2* gene on the X chromosome at Xp22.2 has an allele frequency of 0.36 and an odds ratio of 1.07. This gene acts as a tumour suppressor which contributes to the control of cell motility, and the risk allele is an eQTL which reduces SHROOM2 expression in the colonic mucosa[98].

1.1.8 Environmental risk factors and gene-environment interactions

There are multiple environmental risk factors which are proposed to modulate the likelihood of individuals to develop CRC during the course of their lifetime, some of which are modifiable. A meta-analysis of 106 observational studies concluded a relative risk of 1.25 (95% CI, 1.14-1.37) for mortality from colorectal cancer per person year for ever-smokers vs never-smokers[99]. A meta-analysis of 7 studies of long-term alcohol consumption showed a positive linear dose-response with CRC incidence, with a relative risk of 1.49 (95% CI: 1.27, 1.74) between the lowest and highest intake categories[100].

Red meat consumption is proposed to increase the risk of colorectal cancer due to the production of carcinogenic N-nitroso compounds and aldehydes during the digestion of iron-rich haem, and from the nitrite composition that preserved or processed meats often contain[101]. Nevertheless, the overall relative risk increase between the highest and lowest red meat consumption from meta-analyses was only 1.11 (95% CI: 1.03–1.19), and there was significant heterogeneity in the association between populations and demographics[102]. Conversely, dietary fibre is hypothesised to reduce CRC risk, theoretically by diluting the concentration of any carcinogenic metabolites in the lumen and through the release of short-chain-fatty-

acids upon its digestion by gut microflora which reduce intestinal pH and have anti-inflammatory properties[103,104].

The caveats with all observational and retrospective studies is they may be subject to recall bias by the participants, and the possibility of confounding interactions and co-associations between many factors, for instance diet, obesity and exercise combined. Obesity prior to CRC diagnosis has been associated with poorer survival prospects[105], and a multiethinic prospective study of 982 individuals found a history of exercise for >1 hour per week associated with a decreased odds ratio for the occurrence of adenomatous polyps of 0.67 (95% CI 0.40 - 0.90)[106].

In a 9.5 year follow up study of a >5,000 strong German cohort aged between 50-74 years at baseline, there was a significant association between circulating 25-hydroxy vitamin D level (25-OHD) and cancer mortality between deficient vs sufficient individuals[107]. In a prospective Scottish cohort of 1,598 patients, it was found that postoperative 25-OHD level was associated with survival outcome (hazard ratio 0.68, 95% CI 0.50-0.90) for patients with stage I-III CRC when comparing individuals from the upper and lower vitamin D tertiles[108].

Gene x Environment interactions, whereby the magnitude of effect of exposure to a particular modifier is influenced by genotype of the individual at a particular locus, contribute to the incidence of almost all complex traits, including CRC. The same study which observed an association between postoperative vitamin D levels and CRC survival was also able to model a statistically significant (P=0.008) interaction between CRC mortality and a combination term of the rs11568820 variant of the vitamin D receptor gene and circulating vitamin D level[108]; however a follow up Mendelian Randomisation study did not find a statistically significant association between 25-OHD score and CRC risk (individual-level OR 1.03, CI 0.51-2.07)[109]. Gene x environment interaction associations with CRC risk have also been hypothesized between smoking exposure and variants of the *NAT2* gene responsible for detoxifying carcinogenic aromatic amines[110], and between the protective effect conferred by NSAID usage and variants of the glutathione transferase enzyme MGST1 which is involved in production of inflammation-mediating prostaglandin fatty acids[111,112]

1.1.9 Contribution of inflammation, immunity and the microbiome

There are numerous links between inflammation and CRC. It is suggested that localised damage of mucosa in conditions such as ulcerative colitis may present the opportunity for mutated clonal cells to expand during the epithelial proliferation necessary to repair and initiate new crypts[12]. Or, signals from an inflammatory microenvironment such as cytokines could promote de-differentiation of non-stem cells[41]. Coupled with the fact that inflammatory cells can release increased levels of reactive oxygen species which readily form DNA adducts, cells with somatic mutations which de-differentiate could lead to the generation of cancer stem cells - as would be proposed by the "top-down" hypothesis of tumour initiation[113]. 1-2% of CRC cases are attributed as a direct consequence of inflammatory bowel disease, and 15% of all deaths of IBD sufferers is as a consequence of CRC with risk of developing the disease increasing 0.5-1% yearly following diagnosis[114]. A meta-analysis of 60,122 patients with Crohn's disease found that affected individuals had a relative risk of CRC 28.4 times that of the baseline population (95% C.I. 14.46-55.66)[115]. Whilst it could be that individuals independently co-inherit predisposition to both IBD and CRC, inflammation may be culpable for CRC predisposition because it has been observed that long-term administration of non-steroidal anti-inflammatory drugs can have a protective effect in reducing incidence of CRC. One meta-analysis of 25,570 patients receiving a daily dose of aspirin for ≥5 years (originally for the intention of preventing vascular events) had their odds ratio of CRC reduced to 0.79 (95% C.I. 0·68–0·92)[116]. A retrospective study of 1,594 IBD patients similarly proposed that the anti-inflammatory mesalamine may have a protective effect against development of CRC in this population[117]. Although the nature of the protective mechanism is as yet unconfirmed, there are theories under active investigation that the inhibition of mTOR signalling by such drugs could be the link[118,119].

The microbiome is increasingly being implicated in initiation and progression of cancer, and associations have been found between CRC and presence of certain bacterial populations within the intestine, where 99% of the microorganisms that humans host are located[120]. There may be an additional contribution of diet which is either receptive or unfavourable to commensal or pathogenic microorganisms. Infection with *Helicobacter pylori* has been found to be associated with colorectal neoplasia and cancer[121], presumably due to their secretion of cyctotoxin-associated

gene a (CagA) into host luminal cells, which has inflammatory and carcinogenic properties[122]. The presence of *Fusobacterium* species in colorectal carcinoma tissue associates with a poorer prognosis[123], which could be as a result of the increased expression of growth-promoting interleukin 17 and TNFα which are produced by the gut immune cells in their presence[124]. Certain distinct microbiomes have been detected in individuals with different consensus molecular subtypes of colorectal cancer, with the CMS1 subtype being the most drastically altered compared to other subtypes with individuals having larger relative populations of certain *Fusobacterium* and *Porphyromonas* species[125]. The link between the CMS1 phenotype and the microbiome could be because one of the primary fermentation products of gut bacteria are short chain fatty acids such as butyrate[126], which have been shown to contribute to regulation of cytokines in colonic macrophages by altering the activity of histone deacetylases[127] - similar to the epigenetic dysregulation that is a hallmark of CMS1 tumours.

1.1.10 Heritability and missing heritability

Multiple estimates for the heritability of CRC have been made

In 2000, Lichtenstein *et al.* estimated the relative contribution of heritable and environmental factors to incidence of sporadic cancers using a cohort of 44,788 monozygotic and dizygotic twins from Sweden, Denmark and Finland[128]. Using the concordance of incidence of CRC in genetically identical monozygotic twins or dizygotic twins (assumed to be 50% genetically related), structural equation modelling was used to fit a model to estimate the contribution of narrow sense heritability, shared environmental effects (e.g. childhood diet/exposure to second hand smoke) and non-shared environmental effects (which captures any stochastic incidence of oncogenic mutations) to risk of developing CRC. This study estimated the heritability of CRC to be 35%, however there is a large 95% confidence interval attached to this ranging from 10% to 48%. The range is likely large because of the small number of twins included in the study; there were 1,262 twins discordant for CRC and 62 concordant. The same model estimated shared environmental factors to contribute 5% (95% C.I. 0.00-2.3%) and non-shared environmental factors to contribute 60% (95% C.I. 52-70%). This model indicates that environmental factors are the major contributor to colorectal cancer risk, with a smaller yet significant genetic component. Although there is a higher incidence of CRC in males than in

females, this study found no difference in genetic heritability between genders. Cohorts were also split by age into those developing CRC at younger or older than 75 years (35% of CRC cases occurred in twins before the age of 63). Heritability explained a larger proportion of variance in incidence of CRC in the younger cohort, likely due to an increased probability of sporadic mutations causing cancer in older twins. This study made a number of assumptions including: random mating between parents, identical environments, susceptibility to cancer following an underlying normal distribution, dizygotic twins sharing exactly 50% of their alleles, and no interaction between genetic and environmental components contributing to risk of CRC. Any environmental contribution identified in this study may also be population specific, being based on only Nordic cohorts.

The same group carried out another study in 2002 making use of the entirety of the Swedish Family-Cancer Database to leverage relatedness information for all possible pairs of relations, not just twins[129]. They assumed an underlying normal distribution of liability to each cancer type, with liability being contributed to from multiple different sources: additive genetic factors, shared living environment, shared childhood environment and non-shared environmental factors (which includes sporadic somatic mutations). They estimated the proportion of heritability of each cancer attributable to additive genetic factors based on tetrachoric correlations of incidence between different relations who shared different proportions of different sources of liability. From this structural equation modelling they estimated the additive genetic heritability of CRC to be 13% (95% C.I. 12% to 18%). Given the differences between the environment shared between siblings or between parent-offspring pairs, this study was now able to partition shared-environment into two sub-categories (compared to the single category from the group's previous 2000 study[129]), one being shared childhood environment. This shared childhood environment was calculated to contribute 6% (95% C.I. 5% to 7%) of the heritability of CRC (it was more important in other cancers such as Stomach and Cervix *in situ*, both where it contributed 13%). Regular shared environment, including that shared between generations such as in parent-offspring pairs, contributed 12% (95% C.I. 11% to 13%) to CRC heritability. The single largest factor contributing to CRC liability was still non-environmental effects incorporating stochastic somatic mutations at 69% (95% C.I. 68% to 70%). This study improved upon the group's previous iteration because it was able to use a much larger sample size and deconvolute shared environmental effects into two distinct compartments.

Limitations of this study are that it likely suffers from population-specific biases having only used data from Swedish individuals.

Jiao *et al.* took a different approach in their 2014 study. Instead of using explicit family relationships, they simply analysed genotypes for 8,025 cases and 10,814 controls from SNP arrays and calculated their genetic similarity based on the observed variants. Assuming high-quality genotyping, this technique should be more accurate than relying on self-reported family relationships, which assume relatedness between siblings to be 0.5 when there could be much variability around this depending on meiotic recombination, and which also may suffer from occasional errors of incorrectly attributed parentage. Using a restricted maximum likelihood, they estimated the heritability of CRC to be 7.42%, only 0.65% of which was explained by variants within 250Kbp of the 31 known associated GWAS loci at the time[130]. They estimated a Genotype x Environment contribution of smoking to account for 6.94% of the variance in heritability of CRC, however a limitation of their study was that they only had relatively sparse genotyping data from 550K and 730K CHIPs, with a maximum of 620K with MAF > 0.01.

Muñoz *et al.* assessed CRC heritability from two separate cohorts of UKBioBank individuals in their 2016 study. They first utilised 1.56M records of self-reported family histories of CRC, and heritability was estimated using either a simple family-based model or structural equation modelling which took into account shared family environments. The second was genotypes from 525K SNP arrays for 114,000 unrelated individuals. Their estimates for genetic heritability of CRC was 26% (CI 24%-28%) from family-based models, 24% (CI 21%-26%) from SEM and 12% (CI 0%-28%) from the SNPs[131]. This highlights that shared familial environment likely explains a portion of predisposition to CRC, and that SNP arrays fail to comprehensively capture 100% of the genetic variation responsible for predisposition.

Sources of Missing Heritability

There may be missing heritability from these estimates due to SNP-arrays used for GWAS studies not capturing structural variants, copy number mutations, indels or epigenetic modifications. The estimates reported in these studies only represent narrow-sense heritability not broad-sense, and not all Gene x Environment or Gene x Gene interactions would be captured in these models.

Corradin *et al.* propose another source of potential missing heritability: variants which fall outside of the LD blocks of GWAS SNPs, but through which the 3D architecture of the genome are brought into proximity with GWAS SNPs and interact with them[132]. They find that the genotypes of these proximal variants can influence the magnitude of clinical disease risk ascribable to GWAS SNPs, as well as modulating expression of transcripts relevant to the disease condition.

1.2 Central Dogma: DNA - RNA - Protein

1.2.1 Structure of DNA

Deoxyribonucleic acid stores genetic information in the nucleus of all eukaryotic species using the bases adenine (A), cytosine (C), guanine (G) and thymine (T). The human genome has an average GC content of 46.1%, though it ranges between 35 to 60% with a bias for greater GC content in genes and regulatory regions[133]. Genes are the basic unit of genetic storage which encode the order in which amino acids should be assembled if the gene is destined to produce a polypeptide. The portions of gene sequence which code for amino acids are termed exons, and in eukaryotic genomes they are interspersed by larger tracts of non-coding sequence called introns. Introns are identified by a GU dinucleotide at their 5' end, and an AG at their 3' end. There is also a "branchpoint sequence" (BPS) located within the intron, usually towards the 3' end, which mediates binding of key components of the spliceosome, and following this there is a polypyrimidine tract of up to 50bp which aids in spliceosome recognition of the BPS[134](*Figure 1.4*). Introns are excised co-transcriptionally from nascent mRNA molecules as they are produced by RNA Pol II[135], and they allow for greater flexibility of gene expression by facilitating the production of alternative isoforms from the same original gene sequence.

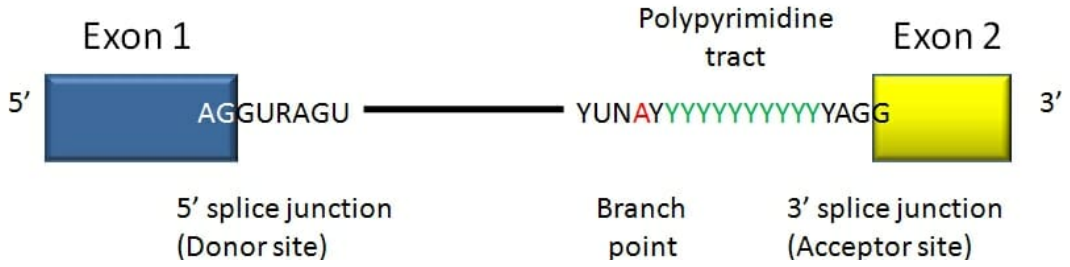


Figure 1.4 Sequences defining splice sites of an intron. From Biologydictionary.net[136].

Promoters and enhancers

Promoters are ~100-200bp sequences located immediately upstream of genes which have affinity for and recruit transcription factors (TFs), potentially aided in their binding by co-factors which do not contact the DNA directly[137]. TFs aid recruitment of members of the basal transcription machinery, including the RNA polymerase complexes, and mediate the accurate initiation of transcription at the appropriate transcription start site (TSS)[138]. Enhancer sequences often lie distal to promoters up or downstream of the genes they control, and as far as >1Mbp in the case of the *SSH* gene and its ZRS enhancer element which controls limb bud formation during development[139]. Enhancer sequences also have affinity for TFs, and provide extra impetus for gene expression by recruiting activator proteins, bringing the transcription machinery into close proximity with the promoter region, and stimulating chromatin remodelling to keep DNA open and accessible[140]. There can be multiple different TSSs for a given gene[141], and the spatiotemporal control of which site is used to what degree in different tissues and developmental stages is determined by the cell-type-specific expression of transcription factors and the chromatin accessibility of necessary enhancer sequences[142]. Clusters of enhancer elements that interact physically have been termed “super-enhancers”, and can play a role in activating genes which are key to defining the identity of a particular cell type[143]. They are regularly bound by TFs that constitute the culmination of a cell signalling pathway, and they have been found to harbour higher densities of non-coding disease associated variants than expected given their raw sequence content (2.4x enrichment, based on a survey of 5,303 SNPs identified from 1,675 GWAS studies for a variety of traits)[144], highlighting their importance for healthy tissue development and homeostasis.

RNA polymerase

The RNA polymerase II complex is responsible for the majority of transcription of active genes in eukaryotic cells[145]. The 550kDa complex is assembled from a minimum of 12 separate proteins, including a pre-initiation complex of transcription factors which are responsible for unwinding the DNA helix, the stabilising mediator complex, and RPB1 which performs the main catalytic function of polymerising RNA bases according to the template DNA strand[146]. RPB1 contains repeats of a highly conserved heptad of amino acids at its C-terminus, which become phosphorylated on serines 2 and 5 in order to activate transcription[147]. The complex has an in-built proofreading capability to cleave-out mis-incorporated nucleotides, which achieves

an error rate of approximately 1 in $1x10^6$ bases[148]. The elongation rate of the RNA Pol II complex ranges between 1 and 6 kbp per minute, with its processivity being dependent on chromatin structure and local epigenetic modifications[149].

1.2.2 The spliceosome and alternative splicing

Splicing is mediated by the spliceosome, a complex of multiple ribonucleoproteins (RNPs), which oversees the necessary two-step process to remove each end of an intron from a sequence of pre-mRNA and fuse the ends of the adjacent exons together (*Figure 1.5*). A core of approximately 80 different RNA and protein components are conserved between yeast and humans, with up to 175 in total contributing to mammalian splicing[150], which highlights the increased variety and level of regulatory control over splicing exhibited in higher eukaryotes[151]. The key ribonuclear proteins necessary for the function of the spliceosome are named U1, U2, U5 and U4/U6 after their uridine-rich RNA components. Each contains a small nuclear ribonucleic acid (snRNA - two in the case of U4/U6) which base-pair with each other and the nascent mRNA, and with a variable number of associated proteins which aid the dynamic conformational changes necessary to produce the catalytic active site of the spliceosome[152]. Unlike the RNA Polymerase II complex, which can transcribe an entire transcript after assembling once, the entire spliceosome needs to re-assemble to excise each intron of a gene sequence. During the process of alternative splicing, the intron-identifying features are recognised multiple times by different components of the spliceosome complex to ensure accurate splicing and to provide opportunities for fine-tuning of regulation. The cleaving of the nascent mRNA backbone is facilitated by a trans-eseterification reaction, which itself is independent of ATP. However ATPases and GTPases are required to mediate the complex conformational rearrangements that are necessary for the spliceosome to transition from one active state to another. These include DExD/H-type ATPases/helicases, which have conserved amino acid motifs of Asp-Glu-any-Asp (DExD) repeats[153], and peptidyl-prolyl cis/trans isomerases (PPIases). The active site of the spliceosome is constructed from the snRNAs, the mRNA itself, and contributions from the U2 complex protein SF3B6 and the helicase PRPF8[154,155].

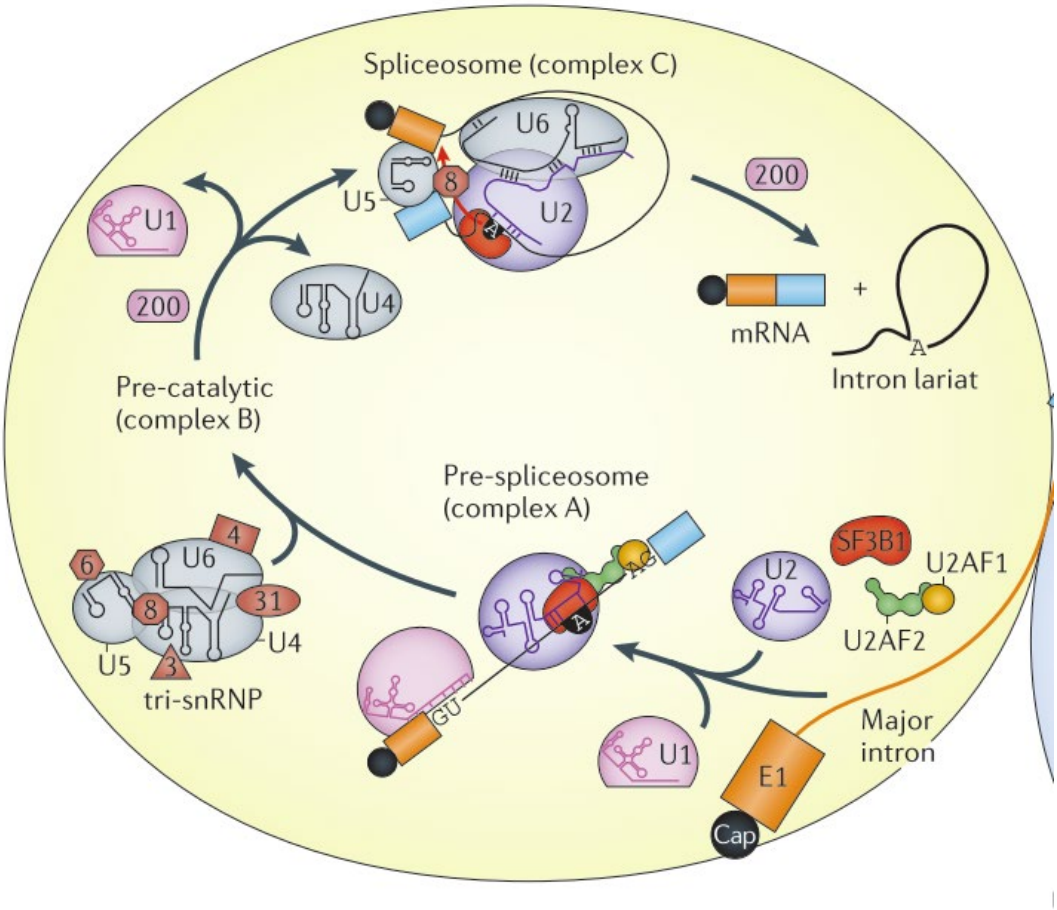


Figure 1.5 Splicing via "Intron-recognition" mechanism. From Scotti *et al.* 2016[156].

Splicing mechanism

The intron recognition mechanism is conserved from yeast to humans. The first step is identification of the 5' SS by the U1 snRNP and its ATP-dependent formation of hydrogen bonds with the sequence, which is facilitated by the DExD-helicase DDX46. The sequence-based interaction is relatively weak by itself, so is stabilised by the protein components of the U1 subunit and serine-arginine-rich proteins (SR) bound to any nearby exonic splice enhancer sites (ESSs)[157]. Next, the SF1 protein selectively binds to the branchpoint sequence[158] and the U2 auxiliary factor (U2AF) binds to the polypyrimidine tract. U2AF has a 65kDa and 35kDa domain, the former of which interacts with the C-terminal RNA recognition motif of SF1 to mediate their coordinated binding with the intron. The 35kDa portion of U2AF simultaneously binds to the 3' SS of the intron. These entities together constitute the spliceosome

“E complex”, also known as the “commitment complex”, which signals commencement of the splicing process.

The “A complex” is then formed as the snRNA component of U2 replaces SF1 on the branchpoint binding site in an ATP-dependent manner. This demonstrates the pattern of the spliceosome requiring multiple different components to recognise the same mRNA elements sequentially. The interaction is stabilised by the SF3a and SF3b protein complexes of the U2 snRNP, and by the arginine-serine-rich domain of the 65kDa U2AF subunit[159]. At the same time, the SF3B14 protein subunit of U2 binds to the final adenosine of the branchpoint site which will take part in the first transesterification reaction to excise the 5’ end of the intron[160].

The next stage of the process is the recruitment of the U4/U6 and U5 snRNPs to form the “B complex”. The U4/U6 and U5 snRNPs associate as the pre-assembled “tri-snRNP” which form in Cajal bodies, microscopically visible membrane-less compartments within the nucleus which are possibly stabilised by means of phase separation mediated by intrinsically disordered domains of their protein constituents[161]. When recruited, the U5 snRNA component makes contact with nucleotides in both the 5’ and 3’ exon, and the U6 snRNA binds to U2[162]. The B complex now contains all of the core snRNPs of the spliceosome, however it is not catalytically active. That requires dissociation of U1 and U4 to form the “B* complex”, a change of state which requires the DExD helicases DDX23 and SNRNP200 under the control of PRPF8 and the GTPase EFTUD2[163,164].

Whilst part of the combined U4/U6 snRNP, the U4 snRNA pair-binds with a highly conserved “ACAGAG” motif of the U6 RNA. During the rearrangement from B to B*, the U6 motif is freed as U4 is displaced so that it may bind to the 5’ SS and the surrounding sequence of the intron[165]. This forms the “C-complex”, whereby the adenosine of the branchpoint sequence is brought into sufficiently close proximity with the guanine of the 5’ SS for its 2’ OH to perform nucleophilic attack on the phosphodiester bond as part of the first transesterification reaction. This cleaves the phosphate backbone at the boundary of the exon and leaves it with a free 3’ OH[152].

Another conformational change then occurs, mediated by DHX38 and DHX8[162], which brings the OH in to contact with the first nucleotide after the guanine of the 3’ SS and allows for the second transesterification event to take place. This produces an mRNA with the ends of the two proximal exons ligated together, and the

spliceosome, still bound to the excised intron lariat, now dissociates in order to be recycled for future catalysis. SNRNP200 is required again in order to aid the disassembly, as is the helicase DHX15[166]. At this stage of the cycle, the U5 snRNP has accumulated protein chaperones which amount to a particle of size 35S. The snRNP must dissociate back down to a size of 20S before it can recombine with the U4/U6 complex to re-form the tri-snRNP necessary to initiate intron catalysis[150].

For the longer introns found commonly in higher eukaryotes, an alternative mechanism termed “exon recognition” can be used to initiate the formation of the spliceosome complex[167]. In this situation, U1 first binds to the 5’ SS downstream of the exon in question, whilst U2 and U2AF bind to the branchpoint sequence and the polypyrimidine tract at the 3’ end of the intron upstream of the exon. SR proteins binding to exonic splice enhancer sequences within the exon mediate cross-exon associations between these two subunits[150]. There subsequently occurs a conformational rearrangement that is not yet characterised, but involves the transition from the cross-exon, A-like complex to the canonical cross-intron B complex, from which intron canonical removal can proceed[168]. RBM5 is a regulatory protein which exploits this process to inhibit the apoptotic signalling protein FAS, because it blocks transition from a cross-exon to cross-intron state and so prevents the inclusion of the gene’s exon 6 which makes the protein membrane soluble and so able to promote programmed cell death[169].

Determinants of splicing

It is clear that many different mRNA sequences have the ability to influence the decision to excise an intron, as do the relative expression levels of a multitude of accessory proteins and transcription factors. Splicing of introns does not simply occur in the order that introns are transcribed - otherwise the facilitation of alternative splicing by competition between splice sites would not be feasible. Introns of the FGA gene were found to be preferentially spliced in the order: intron 3, intron 2, intron 4 and intron 1, and mutations altering the specificity of these splice sites or introducing cryptic splice boundaries can lead to a lack of circulating fibrinogen and the Mendelian disorder congenital afibrinogenemia[170].

Introns whose splice site recognition sequences have the greatest affinity for the U1 and U2AF complexes which initiate the splicing process are those most efficiently excised. However the splicing code is degenerate in higher eukaryotes[151], therefore

other additional signals influence the likelihood of introns to be removed. Exonic splicing enhancer (ESE) and silencer (ESS) sequences bind serine-rich SR proteins and heterogenous nuclear RNPs proteins respectively[171], and in concert with similar positively and negatively influential intronic sequences (ISEs and ISSs) play a significant role in determining the tissue specific patterns of intron excision (*Figure 1.6*)[172]. The expression of different splicing-associated factors can also influence patterns of alternative splicing, e.g. PUF60 has homologous binding sites to U2AF65, but plays a more important role in the recognition of introns with weaker 3' splice sites[173].

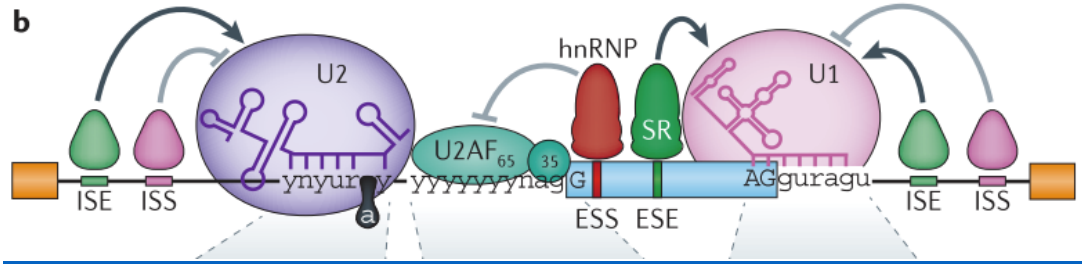


Figure 1.6 Splicing regulatory sequences. From Scotti *et al.* 2016[156].

Secondary structures of mRNA, such as hairpin-loops and G-quadruplexes, may also play a role in regulating the activity of the spliceosome, by obscuring spliceosome recognition elements or by bringing the ends of the intron into closer proximity with the catalytic centres[174]. There exists a number of "class II" introns in the human genome which are able to self-catalyse their own excision, using the same chemistry of transesterification as the matured spliceosome does. They represent ancient sequences which predate the spliceosome and are likely from where its component snRNA sequences originally evolved[175].

The use of different promoters can also influence downstream splicing decisions. This effect is not mediated simply by the strength of the promoter, but by influences of the promoter on either elongation rate of RNA Pol II, or on the particular set of splicing factors recruited to the Pol II C-terminal domain[176]. Using the in vivo reporter system of a minigene containing exon 33 of human fibronectin, small molecule drugs which inhibit elongation rate of Pol II promote more inclusion of the exon[177], as do Pol II mutants with slower processivity[178]. This suggests that a slower elongation rate of Pol II allows more time for recognition of relatively weaker splice sites through assembly of spliceosomal components before stronger splice sites are

transcribed and become available to compete for their binding. Converse cases have also been observed, for instance slower Pol II elongation can allow more time for binding of the negative regulator of splicing ERT-3, which competes with U2AF65 at the polypyrimidine tract, leading to increased skipping of CFTR exon 9[179].

Epigenetics has also recently been demonstrated to play a role in influencing alternative splicing, both by the architecture of chromatin compaction and by means of specific post-translational modifications of histones. The observation has been made that the median length of exons (137bp) is uncannily similar to the length of DNA bound by a single nucleosome (147bp)[152], and that introns have a sparser nucleosome content than exons[180]. It has been hypothesised that introns evolved from insertions of transposable elements, given that the sequences most amenable to these intercalations are the linker regions in between nucleosomes[181]. The occupancy of exons by nucleosomes alters the transcription kinetics of RNA Pol II, as it must pause and wait for the DNA to transiently dissociate from histones before proceeding[182]. By slowing down the rate of transcription, there is increased opportunity for the proper recognition of exon-delineating sequences and so increased likelihood of intron excision[177].

A further mechanism of epigenetic influence may be via the recruitment of specific splicing factors. H3K4me3 is enriched at the 5' ends of genes, and has been shown to assist in recruitment of early-assembling members of the spliceosome complex, such as the U2 snRNP, through interactions with the chromodomain helicase binding protein CDH1[183]. H3K36me3 is a histone modification which is most commonly deposited at nucleosomes colocalised with exons[184], and has been shown to recruit the splicing silencer PTBP1 (polypyrimidine tract binding protein 1) by means of the adaptor protein MRG15, which masks spliceosome recognition sequences and so can promote skipping of the proximal exon[185]. The chromatin-associated protein PSIP1 has also been shown to associate with H3K36me3, and to increase the localisation of the splicing factor SRSF1 to alternatively spliced exons[186]. In the same way that epigenetic profiles influence tissue-specific gene expression, they likely also play a part in orchestrating particular patterns of splicing[187].

1.2.3 mRNA maturation

In addition to splicing, there are further maturation processes which nascent mRNA must go through before exiting the nucleus and entering the cytoplasm. The 5' end of the mRNA molecule is "capped" by the addition of an extra guanosine through a non-standard 5' to 5' phosphodiester bond, which is then methylated at the 7-nitrogen[188]. The cap differentiates mRNAs destined for translation into protein from nucleus-resident RNAs, such as the spliceosome's snRNAs, by facilitating export to the cytoplasm through nuclear pores and accurate entry of the mRNA into the ribosomal translation machinery[189].

The 3' end of any transcripts produced by RNA Pol II (except for a specific set of histone-coding transcripts which are expressed during the replicative S-Phase of the cell cycle) undergo cleavage by the CPSF endonuclease complex, which is directed by polyadenylation recognition motifs that are generally A and U rich[190]. Multiple adenosine mono-phosphates are then added to the 3' cleaved end by poly A polymerase (PAP), which creates a "poly-A tail"[191]. This is a necessary requirement for mature mRNA to be exported from the nucleus, and once in the cytoplasm protects the transcript from premature degradation by endogenous exoribonucleases[192]. Up to 70% of eukaryotic transcripts have multiple possible polyadenylation sites which can produce 3' untranslated regions of different lengths and sequence contents, and some alternative sites are located prior to certain exons, meaning that cleavage can shorten the gene sequences and lead to loss of coding regions in a manner similar to that of alternative splicing[193]. Ribosomal and transfer RNAs are synthesised by RNA Pol I and RNA Pol III respectively and do not possess poly-A tails[194,195].

The 5' and 3' UTRs (untranslated regions) of transcripts can contribute to the control of gene expression in a variety of ways. The 5' region is key for efficient recruitment of the mRNA to the ribosome, and can influence the choice of start codon through affinity for translation initiation factors or by adopting different secondary structures which may mask certain start codons and promote the use of different open reading frames[196]. The 3' UTR can harbour binding sites for miRNAs which induce degradation of isoforms via the RISC complex, and so shorten the half-life of transcripts in the cytoplasm[197]. They can also interact via AU-rich elements with RNA binding proteins (RBP) which can influence mRNA targeting and localisation[198]. Different alternative 3' UTRs of the same transcripts produced as a result of

alternative cleavage and polyadenylation sites can form different secondary structures which could either obscure or present miRNA and RBP binding sites[199].

Classes of alternative splicing events

All of these mechanisms for differential regulation of splicing and mRNA maturation provide the opportunity for great diversity of alternative transcripts to be produced from the same gene. There can be exon-skipping events with one or more exons included or excluded, and there can be mutually exclusive exons whereby the utilisation of one splice site concomitantly influences the inclusion of a different exon (*Figure 1.7*). A well-characterised example of this is exons IIIb and IIIc of the fibroblast growth factor receptor, FGFR, which have tissue specific patterns of mutually exclusive expression in normal prostate epithelium or mesenchymal stem cells[185]. Individual exons can be extended at their 5' or 3' ends by the use of alternative splice sites, and similarly the 5' and 3' UTRs can be lengthened in either direction or alternative UTRs commencing from different points can be used. Usage of an alternative first exon can be triggered by the direction of transcription machinery to different transcription start sites. If a splice site is mutated or masked to such a degree that it is not recognisable at all by the spliceosome, then intron retention can occur. Combinations of multiple alterations can arise within the length of a single transcript, which are termed "complex events"[200]. It has been estimated that up to 94% of human genes undergo alternative splicing, and that in 86% of those events at least one of the isoforms other than the most highly expressed still constitutes at least 15% of the total expression, implying that these alternative transcripts are unlikely to simply be an artefact of imperfect or noisy splicing[201].

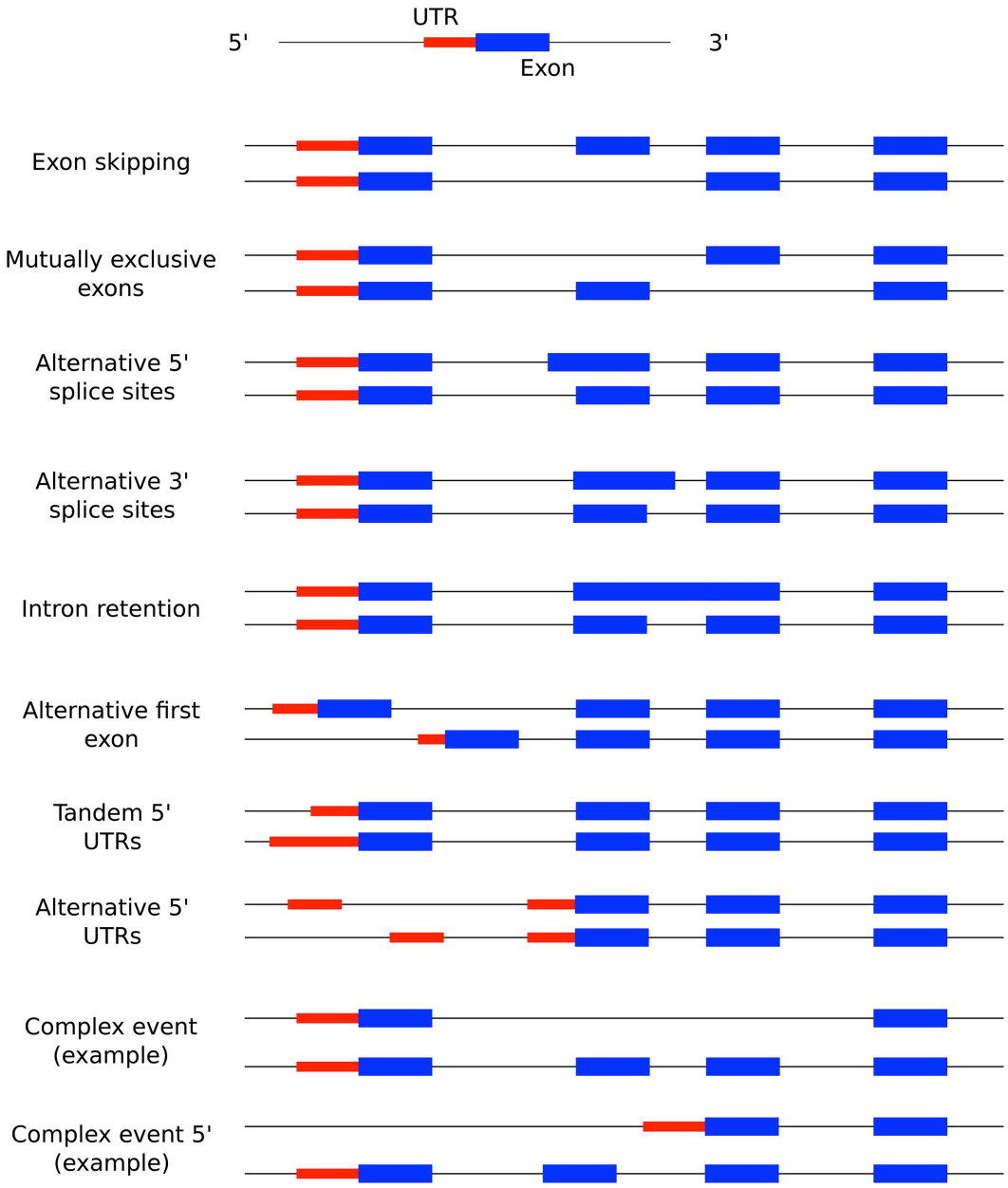


Figure 1.7 Different possibilities and combinations of alternative splicing events. From Ardlie et al. 2015[200].

Diseases caused by mis-splicing

There are well-characterised examples of germline mutations in key splice-site recognition sequences causing severe Mendelian disorders, which highlights the importance of expression of the correct transcript in the correct tissue at the appropriate time. There are very few mutations found in the snRNA components of the snRNP complexes however, and such mutations are likely embryonic lethal[156].

A mutation in an alternative 3' splice site 19bp upstream of exon 2 of the β-globin gene results in a transcript with a premature stop codon and causes β^{+}-thalassaemia[202]. Duchenne and Becker Muscular Dystrophy can be caused by skipping of exons 45-55, or by creation of a novel splice donor and acceptor sites in the DMD gene[203]. A mutation which creates a novel 5' splice site recognisable by the U1 snRNP in exon 7 of the PINK1 gene has been shown to cause early-onset Parkinson's disease from analysis of 31 members of a Spanish family with a historical pedigree of the condition[204]. Multiple disorders of different pathologies can arise as a result of various mutations perturbing different aspects of splicing of the LMNA gene. In healthy individuals, LMNA encodes two separate nuclear lamin peptides from alternative splicing of the same primary transcript, which both serve to maintain nucleus structure. Mutations at the 3' splice site of exon 4 of LMNA lead to extension of the exon by 9 nucleotides, and the addition of the resulting 3 amino acids to the resultant protein leads to dilated cardiomyopathy, in which there is suboptimal development of the heart's ventricles[205]. Intron retention of either intron 8 or 9 of LMNA can lead to type 2 lipodystrophy or limb girdle muscular dystrophy respectively, both of which are caused by the resulting aberrantly spliced transcripts possessing premature stop codons and being degraded by the Nonsense Mediated Decay pathway (NMD)[206,207]. The premature ageing disease Hutchinson-Gilford progeria syndrome can be caused by introduction of a cryptic 5' splice site via mutation within exon 11 of the LMNA gene, which results in a 150bp deletion as the remainder of the exon is not used and is instead joined directly to the 5' end of exon 12[208].

1.3 GWAS

1.3.1 Introduction to GWAS theory and methodology

Genome wide association studies (GWAS) allow for hypothesis-free identification of variants which are significantly associated with disease phenotypes. They require the recruitment of large numbers of diseased and non-diseased individuals for genetic profiling[209]. GWAS is usually performed with genotypes derived from SNP-arrays. This limits the scope of the genome which can be surveiled; however imputation can be used to extrapolate the presence of some variants not covered by arrays based on external databases of human haplotypes such as the 1000 Genomes cohorts[210]. Summary statistics from multiple individual studies can be

combined in a meta-analysis using fixed-effects or random-effects models, depending on assumed heterogeneity of the samples. Whilst studies of family lineages with strong pedigrees of certain diseases are usually the most fruitful resource for discovering rare variants with high effect sizes, GWAS provides the ability to identify larger numbers of more common yet more moderate effect size variants (*Figure 1.8*)[211].

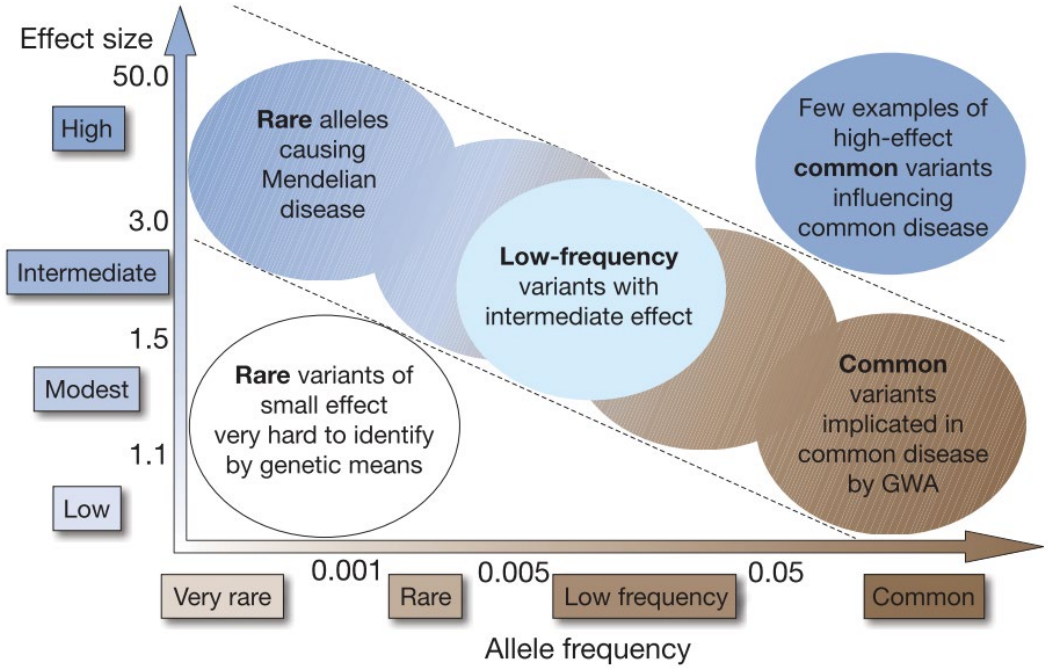


Figure 1.8 Variants amenable to discovery by GWAS as a function of frequency and effect size. From Manolio et al. 2009[211].

Linkage Disequilibrium

Unequal rates of reassortment of genotypes during meiosis in germline cells leads to certain regions of the genome being co-inherited more frequently than would be expected by chance. Variants located in such regions are said to have high linkage disequilibrium (LD) with each other[212]. LD can be quantified by the D' value, which is derived from the expected probabilities of observing each haplotype combination of alleles given the individual allele frequencies observed in the population. D' = 0 when two SNPs are completely independent or in perfect Hardy-Weinberg equilibrium, and values of +1 or -1 mean that there is complete disequilibrium when

one of the four possible combinations of haplotypes between two SNPs is never observed. LD is commonly reported as r^2, which is the square of the Pearson correlation between two haplotypes, calculated by

$$r = \frac{D}{\sqrt{p1 * p2 * q1 * q2}}$$

where p and q are the reference and alternative allele frequencies respectively at each of two loci, denoted by 1 and 2[213]. When using a chi-squared test to assess the deviation of haplotype frequencies from their expected values given the individual allele frequencies, r^2 possesses the convenient property that $\chi^2 = r^2 * N$, where N is the number of chromosomes. Two variants with an r^2 of greater than or equal to 0.8 are generally considered to be in strong linkage disequilibrium with each other[96]. Most SNPs identified by GWAS, especially when using SNP arrays, are likely to not be the causal variant, but instead tag the region of linkage disequilibrium in which the true causal SNP lies[214].

Genomic characteristics of GWAS-linked loci

Up to 90% of SNPs which have been linked to complex traits via GWAS are located in non-coding regions[215–217]. It is theorised that such variants affect predisposition to disease by modulating tissue-specific patterns of gene regulation. Kindt *et al.* assessed the enrichment within 54 genomic feature annotations of 1,909 GWAS-linked variants and any associated SNPs in high LD, using permutation tests which appropriately accounted for intrinsic co-occurrence between certain genomic features (e.g. between gene density, evolutionary conservation and chromatin organisation). Combining all their observations into a general logistic regression model, they found that the top two features which most strongly predicted the presence of a GWAS-associated variant were regions of open chromatin and whether or not the SNP represented an eQTL[214]. They also found a modest, yet statistically significant enrichment of trait-associated risk SNPs within intronic variants, and only a modest enrichment within evolutionarily conserved regions.

1.3.2 eQTLs

eQTLs (expression quantitative trait loci) are variants which are associated with a change in the expression level of a gene, when appropriate confounding variables such as age and gender are controlled for. They can either act in *cis,* whereby the variant is within the same genomic region as the gene whose expression it affects

(commonly defined arbitrarily as ±1Mbp), or in *trans*, meaning that the variant is located further away than this distance or on a different chromosome. eQTLs are observed to be tissue specific, and dependent on environmental conditions at the time of sequencing[218]. eQTL variants are enriched within regulatory features such as transcription factor binding sites at promoters and enhancers, and within regions of open chromatin[219]. Wen *et al.* identified 6,555 cis-eQTLs from lymphoblastoid cells from 420 individuals and found them to be 1.49-fold enriched with the 4% of genome-wide SNPs they predicted to affect binding affinities of TFBSs[220]. eQTLs have also been shown to be enriched within DNase Hypersensitive Sites and enriched for variants linked to complex-diseases via GWAS[219]. Dermitzakis *et al.* analysed RNA-seq and WGS from 4 tissues from the TwinsUK cohort[221], and found that there was a linear association between the likelihood of a lead eQTL SNP being a causative variant and inhabiting a region of DNase accessibility[219]. They also developed an algorithm (CaVEMaN), based on non-parametric bootstrapping, to identify the most likely causal eQTL SNP for a gene when there were multiple significant SNP-to-gene associations. They observed that the candidate causal eQTL SNPs identified with this tool were enriched for having more highly significant p-values from GWAS associations to 16 disease traits when compared to all putative eQTL variants identified in their analysis[219]. Singh *et al*. identified 1,312 eQTLs using array expression from four intestinal tissues of 65 individuals (terminal ileum, ascending colon, sigmoid colon, descending colon), of which 11 were tag SNPs for GWAS associations with Inflammatory Bowel Disease (of which there were 163 identified at time of publication)[222]. Trans-eQTLs, whereby the associated SNP is >1Mb away from the gene in question or is located on a different chromosome, can also be identified in humans. It has been estimated that trans-eQTLs contribute up to 70% of the variance in mRNA expression levels, however individual trans-eQTLs tend to be of lower effect-size and more tissue-specific than cis-eQTLs[223].

Even if a disease phenotype can be explained by a germline coding mutation, non-coding cis-regulatory variants have been shown to significantly influence the penetrance of that variant. Castel *et al.* used individuals from the GTEx consortium (which excludes individuals with inherited disorders) to confirm that in the general population there is evidence of purifying selection against combinations of alleles which would increase the expression of deleterious coding variants. Analysing individuals heterozygous for pairs of regulatory SNPa and deleterious coding

variants affecting the same gene, they found a small reduction (0.70%) of allelic expression of such deleterious coding alleles, but the effect was highly significant when compared to combinations of regulatory SNPs with neutral synonymous coding variants ($P=4.57x10^{-9}$)[224]. They also analysed 615 cancer patients from TCGA, and found an enrichment of haplotype combinations which increased the expression of tumour suppressor alleles harbouring deleterious mutations (relative to a matched set of normal GTEx individuals) or increased the splicing in of exons with similarly disadvantageous variants[224]. The absolute risk of developing breast cancer by age 80 is 54% for women harbouring *BRCA2* mutations, however this increases to 82% if individuals also possess particular polymorphisms in *FGFR2* and *TOX3* which augment the risk by *trans* interactions between the gene products[225].

1.3.3 CRC aetiology explained by GWAS

In addition to the well characterised, high-penetrance germline mutations which strongly predispose to CRC, many lower penetrance but higher frequency variants have been associated with CRC via GWAS[98,226,227]. Such studies have served to demonstrate that CRC follows the same pattern as other complex traits of owing the majority of its heritability to non-coding variants[228], and as the sample sizes in the studies increase, progressively lower effect-size variants are identified[229]. A full list of associations is available from the NHGRI-EBI (National Human Genome Research Institute in collaboration with the European Bioinformatics Institute) GWAS Catalog[230]. The effect size of a variant in predisposing individuals to a disease is commonly reported from GWAS studies as an “odds ratio” (OR). OR is first calculated by partitioning observed individuals into “exposed” or “not exposed” and “healthy” or “diseased”. In the case of a GWAS study, “exposed” relates to an individual possessing a particular allele or not. The OR is the ratio of exposed diseased individuals (D_E) over not exposed diseased individuals (D_N), divided by the equivalent ratio of exposed healthy (H_E) over not exposed healthy individuals (H_N).

$$OR = \frac{D_E/D_N}{H_E/H_N}$$

An OR of greater than 1 implies that the presence of an exposure is correlated with a disease state, though it is not necessarily an indication of whether the exposure is causative.

Variants from non-coding regions can be linked to the genes they are most likely to influence the regulation of through tissue-specific eQTLs, epigenetic annotations (including ATAC-seq) and promoter-capture Hi-C[96,231]. Gene set enrichment analysis of genes linked to CRC predisposition through GWAS has identified pathways well-known to be implicated in dysregulation of mucosal growth such as the Wnt-β-catenin pathway (as highlighted by GWAS associations with the *CTNNB1*, *TCF7L2* and *DACT1* loci), and TGF-β (e.g. *GREM1*, *SMAD7*, *SMAD9*, *BMP2*, *BMP4*)[96,231]. Known tumour suppressors implicated in CRC progression have been identified, including the chromatin remodelling gene *TET2* and the *CDKN2A* and *CDKN2B* cyclin dependent kinase inhibitors located at 9p21.3[231]. New pathways not previously linked to colonic carcinogenesis have also been highlighted by newer, higher-powered meta-analyses, including Krüppel-like factors (*KLF2*, *KLF5*) which promote intestinal stem cell proliferation and endothelial cell blood vessel formation, and members of the Hedgehog signalling pathway (*BOC*, *HHIP*) which encode a Hedgehog co-receptor molecule and inhibitor respectively[231]. Immune-related pathways are another recurrent source of associations, particularly in the MHC region of chromosome 6p21.33, with the genes *HLA-C* and *HLA-DRB5*, *HLA-DRB1* and *HLA-DQA1* all implicated[96].

The regulation of lncRNAs has also been affiliated with CRC predisposition via GWAS. Locus 7p13 contains the gene *SNHG15* which produces a lncRNA that has been shown to bind to the zing-finger domain of SNAI2 and prevent its degradation by ubiquitination, which leads to promotion of the epithelial to mesenchymal transition in colorectal cancer[232]. 17p12 harbours a candidate causal variant in exon1 of *LINC00675*, which is usually downregulated in colorectal cancer tissue compared to normal mucosa because it functions to suppress cell proliferation by modulating Wnt/β-catenin signalling[233]. The locus 9p21 overlaps a super-enhancer located intronically within the antisense lncRNA *ANRIL*, which is observed to be upregulated in CRC compared to matched non-neoplastic tissue, and which leads to reduced cell proliferation and rates of lymphatic metastasis when knocked down in CRC cell lines and mouse models of the disease respectively[234,235]. Law *et al*. identified an eQTL for the lncRNA *RP11–378A13.1*[96], which inhabits the same locus 2q35 in which the variant rs992157 has previously been associated with an OR for CRC risk of 1.10[236]. This transcript has not yet been linked to CRC, but it has been shown to be in the top 20 most differentially expressed genes between lung adenocarcinoma and adjacent non-tumour tissue[237].

In 2018, a GWAS for CRC risk was carried out which made use of shallow WGS (3.8-8.6x) of 1,439 CRC sufferers and 720 controls of European ancestry to increase the coverage of rare variants imputable into larger meta-GWAS studies (eventually incorporating 125,478 individuals). This led to the identification of the first rare variants (MAF < 0.01) implicated in CRC, as well as 39 newly identified common variants and re-validation of existing common loci. The rarest variant to be significantly associated was rs145364999 (MAF 0.0031), located intronically to *CHD1* on 5q21.1. Interestingly, the rarer allele is actually protective, with an odds ratio of reduced predisposition to CRC of 0.52 (95% CI 0.40–0.68)[231], and the authors hypothesise that is might mediate its effects by lowering *CHD1* expression which is required for the NF-κβ signalling necessary in tumour cells whose growth is driven by *PTEN* inactivation[238].

An association at 19q13.33 encompassing the *FUT2* gene (fucosyltransferase II) could implicate gene-environment interactions with the microbiome in predisposition to CRC[96]. Variations at this locus which cause people to be non-secretors of FUT2 at the mucosal surface have been associated with reduced diversity of microbial populations, due to the resultant lack of modification of cell surface glycans which commensal bacteria harness, and have also been linked to Inflammatory Bowel Disease which is known to be a significant risk factor for the development of CRC[239,240].

In 2009, after the first 10 common variants associated with predisposition had been identified, Tenesa *et al.* constructed a model which predicted that approximately 170 common SNPs could account for almost all of the genetic variance for CRC risk[228]. The number of loci identified now stands at >100, however views have changed in that the tail of the “L-shaped” distribution of predisposition probably extends far longer than 170 variants, and will include many trans and epistatic interactions, in addition to variants which have variable penetrance depending on environmental exposures[241].

1.3.4 Polygenic risk scores

One goal of GWAS is the identification of multiple genetic loci predisposing to CRC for combination into a polygenic risk score (PRS) which can identify patients where increased screening efforts should be targeted. The benefit of screening for increasing overall survival has already been demonstrated both in the general

population and in those at high risk of developing CRC: a meta-analysis by the Cochrane Institute of 320,000 participants randomised to either receive screening for occult faecal blood or no screening showed a 16% reduction in relative risk of mortality from CRC[242], whilst a 15 year study of 205 families with a history of Lynch Syndrome (including 745 carriers of MLH1 or MSH2 mutations) showed a decreased risk of developing CRC with surveillance frequencies of 1-2 years compared to 2-3 years[243]. Huyghe *et al.* derived a PRS from 95 SNPs independently associated with CRC, and modelled the age at which individuals at various quantiles along the risk spectrum would benefit from screening. They estimated that >50% of males and 10% of females would benefit from beginning screening at earlier than age 50, which is the current guideline implemented by the NHS in Scotland[231]. One complication of implementing such strategies is that risk allele frequencies and LD structures between true causal and tagging variants vary across populations, meaning that PRS scores currently derived from cohorts with mainly European ancestry will be less applicable to other populations[244].

However, a recent study of 5,675 individuals diagnosed with CRC in Scotland by the CCG Group failed to identify any significant correlations between PRS constructed from common variants and clinically relevant endpoints such as overall survival[245]. Incorporating PRS into well-established predictive models of survival including age, gender and stage at diagnosis also failed to improve predictions[245]. This implies that the genetic components which influence predisposition may be superseded by somatic aberrations which then dictate survival and response to clinical interventions once tumours have developed.

1.4 Splice-QTLs

Whilst half of the GWAS loci identified in a recent meta-GWAS from 2019 did have tissue-specific eQTLs for at least one gene in their vicinity[96], this leaves open the possibility for the remaining 50% to be influenced by a different mechanism of regulatory control, such as by splice-QTLs (sQTLs). sQTLs are variants associated with a change in the relative abundances of different transcripts expressed from the same gene[246] (*Figure 1.9*). They do not necessarily alter the total expression of a gene and therefore can occur independently of an eQTL, though some variants may have the effect of changing both the total expression and relative abundances of

transcripts of a single gene[247]. Like eQTLs, sQTLs can either act in *cis* or *trans*[248], with *cis* variants predicted to disrupt sequences defining or aiding the recognition of splice sites and *trans* being attributable to changes in expression or regulation of core spliceosome components.

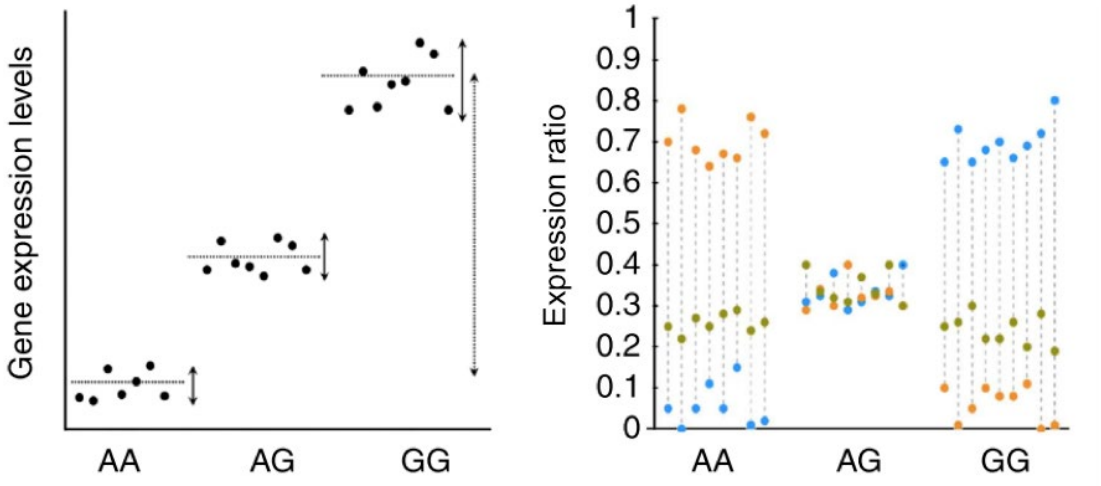


Figure 1.9 Differences between eQTLs and sQTLs
Left panel shows quantification of an eQTL event whereby the total expression of a gene changes in response to a variant. Right panel illustrates an sQTL for a gene expressing three different transcripts denoted by orange, green and blue points. As the genotype changes, there is a concomitant shift in the relative expression or the orange and the blue transcripts, as evidenced by the ratio of total gene expression which they contribute, whilst the green transcript remains unchanged. From Monlong *et al.* 2014[247].

1.4.1 Identifying and quantifying sQTLs

There are a variety of methods for quantifying the phenotype of alternative splicing, which is more complex to capture than total gene expression levels. Techniques have been developed based on the usage of exons, transcripts or introns. One of the earliest studies first calling QTLs from RNA-seq as opposed to microarrays used linear regression to correlate the genotypes of SNPs with the expression of individual exons[249]. They termed their discoveries “sQTLs”, but they could more accurately be called exon-expression-QTLs (eeQTLs). A limitation of this approach is that it wastes power by necessitating many non-independent tests of all individual exons, and it doesn’t preclude identifying classic gene-level eQTLs which would manifest as eeQTLs for every exon expressed by the gene.

A contemporaneous study published in the same issue of Nature by Pickrell *et al.* used “percent spliced in” (PSI) to discover sQTLs. They divided the number of reads

mapped to each exon by the total number mapped to the gene as a whole in order to calculate a ratio of the splicing in of each exon[250]. This has the advantage of controlling for total gene expression, but still requires multiple independent tests for every exon.

Other later studies adopted PSI, but instead of taking each individual exon over the total expression for the gene, they defined PSI for each individual exon as the ratio of reads providing evidence for its inclusion in relation to those providing evidence of its exclusion. This approach was used by the authors of the GLiMMPS package, which only used reads spanning exon boundaries as proof of inclusion or exclusion[251]. They employed a general linear mixed model to account for within-sample uncertainty in PSI as a random effect, and within genotype variability as a fixed effect. Jia *et al.* also used PSI, but incorporated all reads into their estimations of exon-usage, not just those falling on exon boundaries[252]. They adopted a random effects meta regression in order to simultaneously model within-sample uncertainty of exon-level quantification and between-sample uncertainty of expression in relation to genotype (*Table 1.1*).

An inherent limitation of individual exon-centric approaches is that they provide little information about changes at the level of whole isoforms, and therefore can be more challenging to interpret biologically. The first studies to take this approach identified transcript-ratio QTLs (trQTLs), whereby they searched for variants associated with changes in relative expression of whole transcripts in relation to total gene expression[253,254]. Whilst providing more biologically interpretable results, these studies still suffered from the limitation of undertaking multiple non-independent tests for each transcript of a gene.

Therefore multivariate approaches have been adopted more recently to account for the interdependence of all isoforms of a gene, in that the ratios of expression of all isoforms is constrained to always sum to 1.0. The sQTLseekeR package uses the principle of non-parametric multivariate analysis of variance (MANOVA) to assess associations between genotypes and transcript expression ratios[247]. For each genotype of a bi-allelic SNP (e.g. AA, AT, TT), each individual is represented as a point in multidimensional space with as many axes as there are expressed transcripts, where the expression ratio of each transcript provides the coordinate in each dimension. The variance of transcript expression for each genotype is

quantified via a Hellinger distance[255], calculated as the sum of the squared differences between each point and the “centroid” - i.e. the point exactly central to the points from all individuals of that genotype. A non-parametric test devised by Anderson[256] based on MANOVA is used to assess whether the variance between genotype groups is greater than the variance within them. The magnitude of differences between the intra and inter-genotype variability of transcript expression ratios produces an F-statistic, which is assigned nominal significance by comparison to a null distribution constructed from permutations of individuals between genotype groups. Multiple testing correction is then applied via Storey’s q-value to find genome-wide significant sQTLs.

sQTLseekeR also includes a mechanism to control for false positive sQTLs which can be caused by “splicing variance quantitative trait loci” (svQTLs). These are SNPs which correspond to changes in the variance of expression of a feature[257]. A SNP which increases variance in transcript-level expression has clear potential to produce a false positive sQTL - especially if the number of samples corresponding to the allele which causes the greatest variance is relatively small. Therefore sQTLseekeR also calculates the likelihood of each SNP representing an svQTL and discounts such events so that the final list only contains sQTLs where the transcript expression ratios change significantly, but the variance of transcript expression doesn’t.

The DRIMSeq package also takes a multivariate approach, and uses the Dirichlet multinomial distribution to simultaneously model the expression of all isoforms of a gene[246]. An advantage is that it allows the absolute expression levels of each transcript to be taken into account when estimating the uncertainty in relative transcript expression ratios. However DRIMSeq has been shown to be more conservative than sQTLseekeR in indentifying sQTL events when both sQTLseekeR and DRIMSeq were run by the DRIMSeq authors on the same dataset of transcript quantification[246]. Neither sQTLseekeR not DRIMSeq have the ability to apply covariates in their model to capture variation in transcript expression ratio in relation to potential confounding factors, such as age or gender.

The limitation of isoform-centric packages, such as sQTLseekeR and DRIMSeq, is that they require provision of a known transcriptome build and so can only find splicing events involving switching between previously identified isoforms. Altrans is

an exon-based algorithm which uses split reads and read pairs from genomic alignments to quantify the fraction of reads supporting “exon link coverage” between pairs of exons[258]. Altrans then converts the exon link coverages into the equivalent of a PSI phenotype by calculating the fraction of coverage between each pair of exons compared to the total coverage of the gene. Altrans defines exon-boundaries according to a predefined genome build, however it has the notable ability to introduce a degree of novelty by identifying links between exons for which there is no previous record of an isoform possessing such a combination of exons. Altrans requires a separate QTL-mapping algorithm in order to find associations between variants and changes in these exon links. The Altrans authors recommend the algorithm FastQTL, which calculates linear regressions between phenotypes and genotypes in *cis*, and is able to account for any covariates supplied in numeric or ordinal format[259]. FastQTL calculates nominal p-values of association by performing random permutations of the genotypes associated with each phenotype, which can subsequently be corrected for genome-wide multiple testing.

Leafcutter is another PSI-based tool, though it infers the presence of introns as opposed to exon-links. The rationale is that the presence of introns can be captured unambiguously in split reads or pairs of reads further separated than would be expected given the library’s insert size distribution, whereas inferring the structure of an exon which is not entirely spanned by a single read or read pair carries some inherent probabilistic uncertainty[260]. Like Altrans, Leafcutter begins from a bam file of genomic alignments, but it allows for additional novelty because it identifies putative intron excision events *de novo,* agnostically of any genome-build using just the reads observed in the study. Leafcutter constructs graph-representations of splicing events by linking all inferred introns which share at least one intron boundary into a network termed an “intron-cluster”. The equivalent of PSI phenotypes are then calculated as the ratio of usage of one potential intron compared to all other introns belonging to the same cluster. Singleton nodes in the graphs (i.e. intron clusters only containing a single intron) are discarded as these imply splice sites used constitutively across the study cohort and therefore are not informative for sQTL analyses. Once the intronic phenotypes are quantified, Leafcutter again requires an additional QTL-mapping algorithm to find associations, and FastQTL can be applied in this context also. A limitation of Leafcutter is that it is not proficient in identifying alternative usage of 5’ or 3’ UTRs, because these are not usually marked by the presence or absence of introns.

The two algorithms sQTLseekeR and Leafcutter (along with the QTL associating algorithm FastQTL) were chosen to be used in this study as a result of their complementary mechanisms (*Table 1.1*).

Method	Authors	Methods	Dataset	No. sQTLs (FDR level)
Exon expression QTL	Montgomery *et al*. 2010	Variants associated by linear regression with expression levels of exons.	60 CEU LCLs	293 (0.01)
PSI	Pickrell *et al.* 2010	PSI defined as the fraction of reads mapping to each exon divided by all reads mapping to a gene.	69 YRI LCLs	187 (0.1)
GLiMMPS	Zhao *et al.* 2013	Uses only the reads spanning exon boundaries which either support evidence of inclusion or exclusion of an exon. Applies a general linear mixed model accounting for within-sample uncertainty in PSI as a random effect, and within genotype variability as a fixed effect.	41 CEU LCLs	140 (0.1)
Random effects meta-regression	Jia *et al*. 2014	Random effects meta-regression which models within-sample uncertainty of exon-level quantification, and between-sample uncertainty of expression in relation to genotype.	78 CEU LCLs	447 (0.05)
Altrans	Ongen *et al*. 2015	Uses linear regression performed by FastQTL to associate SNPs with changes in the PSI of exon-junction-coverage based on known exon boundaries with potentially novel exon combinations.	373 EUR LCLs	1,427 (0.01)
			89 YRI LCLs	166 (0.01)
Leafcutter	Li *et al*. 2018	Uses linear regression performed by FastQTL to associate SNPs with changes in the PSI of *de novo* inferred intron inclusion.	372 EUR LCLs	5,774 (0.05)
			85 YRI LCLs	1,982 (0.05) 1,294 (0.01)
trQTLs	Lappalainen *et al*. 2013	Calculates the transcript expression ratio for each transcript over the total gene expression, then runs linear associations between each individual transcript and SNP in *cis*.	373 EUR LCLs	620 (0.05)
			89 YRI LCLs	83 (0.05)
	Battle *et al*. 2014	Calculates the transcript expression ratio for each transcript over the total gene expression, then runs non-parametric Spearman rank	922 individuals from the DGN	1,370 (0.05)

		correlations with SNPs in *cis.*	Cohort	
sQTLseekeR	Monlong *et al*. 2014	Uses non-parametric multivariate analysis of variance (MANOVA) to associate SNPs to changes in relative isoform expression.	91 CEU LCLs	155 (0.05)
			95 FIN LCLs	184 (0.05)
			94 GBR LCLs	175 (0.05)
			93 TSI LCLs	185 (0.05)
			89 YRI LCLs	168 (0.05)
DRIMSeq	Nowicka *et al*. 2016	Models isoform expression using the Dirichlet-multinomial distribution, which allows for the level of expression to be taken into account when quantifying uncertainty in expression estimates.	91 CEU LCLs	3,036 (0.05)
			89 YRI LCLs	1,867 (0.05)

Table 1.1 Methods for identifying sQTLs. Abbreviations of GEAUVADIS populations: Utah with European ancestry (CEU), Finns (FIN), British (GBR),Toscani (TSI) and Yoruba (YRI), CEU+FIN+GBR+TSI (EUR). DGN: Depression Genes and Networks[253]. LCLs lymphoblastoid cell lines

1.5 Alternative splicing in complex trait predisposition and cancer

1.5.1 sQTLs in complex trait predisposition

sQTLs have been demonstrated to contribute to predisposition to a number of complex traits.

In their 2016 study, Li *et al*. used Leafcutter to identify 2,893 sQTLs from 86 YRI GEAUVADIS LCLs (note, this is distinct from the 2018 study detailed in *Table 1.1*). Hypothesising that tissue-specific diseases for these white blood cells could include inflammatory and autoimmune disorders, they found the sQTL SNPs to be more significantly enriched than eQTLs in GWAS variants predisposing to multiple sclerosis, and equally as enriched as eQTLs in variants predisposing to rheumatoid arthritis[261].

The same group revisited the topic of sQTLs in disease in their 2018 paper, this time using 372 CEU LCL samples from GEAUVADIS. They identified 5,774 sQTLs at FDR 0.05 and again observed the same pattern of sQTLs being more significantly enriched than eQTLs in variants linked to multiple sclerosis by GWAS, and sQTLs and eQTLs demonstrating parity of enrichment for rheumatoid arthritis[260]. The authors then used S-PrediXcan[262] to perform a transcriptome-wide association study (TWAS) between splicing traits and variants associated by GWAS to 40 complex traits. Supplementing the genetic information with splicing phenotypes allowed them to associate 18 novel genes with rheumatoid arthritis, of which 13 were not able to be identified through the use of gene expression phenotypes alone. Across the 40 traits analysed, inclusion of the splicing phenotypes associated with sQTLs increased the number of genes able to be putatively associated with complex traits by an average of 2.1-fold over using genotypes alone[260].

Lehmann *et al.* combined tumour and normal tissue expression from 282 kidney renal clear cell carcinoma (KIRC) patients, alongside 458,266 germline variants and 128 recurrent somatic mutations to call 228 *cis* and 23 *trans* sQTLs at FDR 0.05. None of the somatic mutations yielded any *cis*-sQTLs, only *trans*-effects. They found *cis*-sQTLs from loci which were annotated as being either causative of or in high LD with variants from the NHGRI-GWAS catalog associated with a range of cancers. An

sQTL in the mitochondrial solute carrier *SLC25* originated from a locus linked to testicular cancer susceptibility, and an sQTL in the *BABAM1* gene has been linked by GWAS to oesophageal squamous cell carcinoma predisposition[248]. None of the variants were associated with KIRC predisposition specifically, though this could be because at time of publication it had received less attention in terms of large multi-centre GWAS studies. There were only 6 known loci associated with KIRC at the time; the most recent meta-analysis in 2017 using 10,784 cases and 20,406 controls has increased the number to 13[263].

Takata *et al*. analysed RNA-seq from post-mortem dorsolateral prefrontal cortex brain tissue of 206 individuals from the CommonMind Consortium. They identified 1,595 significant sQTL events using the PSI method, taking the fraction of read counts supporting each exon's inclusion against exclusion. They tested for enrichment of these sQTLs in SNPs associated with 15 different complex traits from the GWAS catalog when compared with a null population of 48,000 non-sQTL SNPs with matched minor allele frequencies. Of the three disease traits for which there were significant enrichments after Bonferroni multiple testing correction, the greatest effect size was in schizophrenia (OR=3.72, $P=9.9x10^{5}$, one-tailed Fisher's exact text), the others being multiple sclerosis (OR=3.71, P=0.036) and inflammatory bowel disease (OR=1.43, P=0.0014)[264]. The magnitude of significance of the enrichment for the brain-related disease serves to highlight the tissue-specificity of splicing in this organ.

Raj *et al.* used Leafcutter to identify sQTLs in 3,006 genes from post-mortem dorsolateral prefrontal cortex (DLPFC) samples from 450 subjects with a mean age of 88, to investigate the contribution of alternative splicing in the ageing brain to Alzheimer's disease predisposition. The sQTLs they identified from DLPFC were more likely to be enriched for SNPs linked to Alzheimer's via GWAS than other types of QTLs including eQTLs, methylation-QTLs and sQTLs identified from monocytes[265]. Their analysis was able to confirm that alternative splicing was the likely mechanism behind 3 previously annotated GWAS predisposition loci for the *PICALM*, *CLU* and *PTK2B* genes[265], whilst other studies had previously implicated an increase in splice isoforms including the 2nd exon of *CD33* encoding an Ig V-set domain with increased Alzheimer's risk[266,267]. The authors also performed two separate TWASs, one combining expression of 4,746 genes with the germline variants, and one using PSI of 15,013 introns alongside SNPs. They identified more

genome wide associations using the splicing as opposed to expression phenotypes (16 vs 5), of which 8 were associations with novel genes. Performing network analysis of protein-protein interactions using the new and known genes involved in susceptibility, they found a significant enrichment for the endocytosis and autophagy-lysosomal pathways, which makes biological sense given the link between Alzheimer's and dysregulated protein-degradation and trafficking[268]. This demonstrates the benefit of including splicing phenotypes to provide mechanistic insights when investigating predisposition to complex traits[269].

1.5.2 Mutations in spliceosome components can represent *trans*-acting driver events in cancer

Observations of recurrent somatic variants in members of the spliceosome complex have led to them being classified as oncogenes and tumour suppressors. Their mutation leads to widespread dysregulation of splicing, and *trans* effects on many genes.

Seiler *et al*. analysed >10,000 samples from 33 different tumour types from TCGA and identified likely driver genes or tumour suppressors using two complementary algorithms: MutSigCV[270] and a ratiometric approach devised by Vogelstein *et al*.[271], whereby oncogenes are defined as genes showing patterns of repeated mutation at the same codons, whilst tumour suppressors suffer recurrent loss of function mutations along the length of their sequence. The authors detected 119 members from a catalog of 404 spliceosome-associated genes as being significantly enriched for non-silent mutations, with the most recurrently mutated genes tending to be involved in earlier stages of spliceosome assembly - particularly from Complex A though to Complex C[272]. The gene most prolifically affected by hotspot mutations was *SF3B1*, which also showed cancer-type specific patterns of mutation: the R625C/H mutation was most common in skin and uveal melanomas, K700E most frequent in breast cancer, and E902L was specifically present only in bladder cancer[272]. The disease-specific nature of these mutations implies that the particular dysregulations of transcriptional systems caused by each mutation are favourable to survival in different tissue microenvironments. The splicing factor *SRSF2* possessed a recurrent hotspot mutation at proline 95, which has been demonstrated to alter its RNA binding kinetics. This alteration causes a *trans* mis-splicing of the methyltransferase *EZH2* in haematological malignancies, causing its subsequent degradation[273]. Point mutations of *EZH2* itself are common in such blood cancers,

however mutation of the *SRSF2* proline 95 is observed to be almost entirely mutually exclusive with such variants, reinforcing the status of the *SRSF2* mutation as a driver which can produce a similar phenotype to the *EZH2*-specific changes[272]. The splicing-associated RNA binding protein *RBM10* showed patterns of loss-of-function mutations consistent with a tumour suppressive role, which was supported by *in vitro* and *in vivo* experiments where its knockout lead to increased colony formation or tumour growth respectively[274,275]. Whilst mutations in spliceosome factors are likely to have wide-ranging effects on the transcriptional landscape which aid cancer progression, *RBM10* mutation has been specifically shown to cause knock-on mis-splicing of the *TSC2* tumour suppressor in lung cancer adenocarcinoma, which plays a key role in mediating the mTOR pathway in this malignancy[272], and to cause inclusion of exon 9 of the *NUMB* gene which ablates its role in negative regulation of NOTCH signaling[275].

Whilst this thesis will focus on variants influencing regulatory control of alternative splicing in *cis*, the phenomenon of core spliceosomal components being regularly somatically mutated serves the highlight the relevance of splicing regulation and perturbation to cancer.

1.5.3 Aberrant somatic splicing in cancer

Multiple studies have observed perturbations to splicing in cancer which have been implicated in driving progression of the disease

Supek *et al.* searched for evidence of synonymous mutations affecting cancer progression using whole exome sequencing (WES) from 3,851 cancer samples from 11 tumour types. They compared rates of synonymous and missense mutations in known oncogenes and tumour suppressor genes to non-cancer-related genes matched for a variety of molecular characteristics including regional point mutation rates, mRNA expression, histone mark occupancy, base composition and replication timing. They found that oncogenes contained a 23-30% excess of synonymous mutations compared to the matched gene sets ($P=3.0x10^{-6}$), that the synonymous mutations clustered in conserved genic regions similar to patterns of missense mutations ($P=1.0x10^{-6}$), and that they preferentially led to the gain of ESEs and the loss of ESSs ($P=0.02$)[276]. If an oncogene in a particular sample contained one such synonymous mutation, it was then less likely to also contain a driver missense mutation, implying that these synonymous mutations can adopt the mantle of driving

cancers by perturbing spliceosome recognition sequences. Using RNA-seq from 2,000 of the samples and a multivariate distance-based metric to assess differences in the expression and variance of genes' constituent exons, the authors estimated that approximately 50% of all recurrent synonymous mutations had an effect on the spliced content of the oncogenes they inhabited ($P=4.0x10^{-4}$)[277]. The authors estimated that between 20-50% of all synonymous mutations in driver genes had been selected for, making up 6-8% of the total somatic mutations observed in these genes. A corollary of their study was that dosage sensitive oncogenes, i.e. those which most often undergo copy number amplifications in cancer as opposed to SNVs, have an enrichment of somatic mutations in their 3' UTRs (P=0.03), regions which are known to regulate intracellular mRNA levels, thus highlighting a further mechanism by which non-coding somatic mutations in regions regulating mRNA processing can influence cancer[276].

Puente *et al*. analysed 452 cases of chronic lymphocytic leukaemia (CLL) and 54 cases of a precursor disease state, monoclonal B-lymphocytosis (MBL), from the ICGC consortium, and identified recurrent non-coding mutations which appeared to influence disease progression. They observed recurring non-coding mutations in the 3' UTR of *NOTCH1* in 13 patients. These alterations created a new splicing acceptor site within the 3'UTR, which caused the use of a cryptic donor site within the last exon of *NOTCH1* and truncated the last 158 coding bases[278]. The splicing event was confirmed by RNA-seq and western blot, and removed a PEST domain which usually functions to increase protein stability and so ensure proper activity of NOTCH1. Patients with such splice-altering mutations had similar or worse prognoses in terms of time-to-first-treatment and overall survival as patients with previously identified coding mutations in *NOTCH1*, demonstrating the efficacy of this non-coding alteration to drive disease progression.

The Lehmann *et al.* study of KIRC tumour and normal expression identified 16 sQTLs with >25% effect size for genes which were annotated in COSMIC as being involved in cancer progression. Examples included splicing events in the gene *PMF1* implicated in bladder carcinoma, the tumour necrosis factor-associated *C1QTNF3*, and the gene *TMEM176* which was found to express a tumour-specific splice isoform and has been previously linked to lymphoma and lung carcinoma[248,279].

Climente-González *et al.* found that mutations in oncogenes affecting splicing can have the same effects on disrupting protein-protein interaction networks as missense point-mutations. They observed 76 protein domain families which had somatic mutations linked to splicing changes more commonly in driver genes possessing mutational hotspots than non-drivers ($<2.2e^{-16}$, Wilcoxon). Cancer drivers were enriched for splicing changes affecting the gain or loss of a functional protein domain (OR=1.9, $P=2.0e^{-5}$), and these changes occurred independently of driver point mutations in a significant proportion of samples, indicating they could be functioning to promote similar oncogenic processes[280]. An example is the RAC1B isoform of RAC1, which harbours an extra 19 residues downstream of the "switch II" domain which impairs its GTPase activity. This isoform has been shown to be upregulated in colorectal, breast and lung adenocarcinoma, where it synergises with *KRAS* mutations to concomitantly increase tumour cell proliferation[281].

Intron-retention is a particularly potent alternative splicing mechanism through which tumour suppressors can be inactivated in cancer. Jung *et al.* analysed WES and RNA-seq from 1,812 patients of 6 of the most common tumour types from TCGA to identify 1,112 somatic mutations which effectively acted as sQTLs, causing either intron retention (338), exon skipping (503), creation of a cryptic intronic splice site (191) or creation of a cryptic exonic splice site (80)[282]. To ensure that strictly only *cis*-acting SNVs were identified, the authors excluded 62 cancer samples with mutations in genes known to have *trans*-acting effects on splicing, including *SF3B1* and *U2AF1*. The intron retention events affected a subset of genes which were significantly enriched in four different databases of tumour suppressor genes ($P \leq 8.0x10^{-5}$), but not enriched in either of two curated lists of oncogenes ($P>0.05$). 97% of intron-retention events generated a premature termination codon (PTC), with the subsequent mechanism of TSG inactivation likely being nonsense mediated decay (NMD) of the resultant mRNA. From analysis of tumour suppressors commonly identified in their dataset, including *CDH1* and *PTEN*, the authors postulate that intron-retention may act in concert with other mechanisms of tumour suppressor silencing, such as loss of heterozygosity or allele-specific promoter methylation, when individuals are heterzygous for the SNV causing intron retention. *TP53* was the gene in which variants causing intron-retention were most commonly found, with 252 of 1,812 samples harbouring such a mutation[282]. This highlights the pervasiveness of somatic splicing aberrations being co-opted by cancer to facilitate disease progression.

A study by Dvinge and Bradley of 805 matched tumour and normal samples from a more diverse range of 16 TCGA cancers further highlights the importance of intron-retention events. They found evidence for significantly increased rates of intron retention in tumour compared to normal tissue RNA-seq in all except breast cancer (though this could be because of inefficient intron removal in the normal breast controls: an increased intron retention was seen in breast normal tissue vs 14 other normal tissues, but no significant different was seen between breast cancer and other cancers), even in the absence of common spliceosome mutations in components such as *SF3B1, SRSF2* and *U2AF1*[283].

Shiraishi *et al* performed a larger scale analysis of *cis*-regulatory splicing associated variants (SAVs), using 8,976 samples from 31 cancer types with matched tumour/normal RNA-seq from TCGA. They used split reads and exon junction-spanning reads to identify novel intron-exon boundaries and combined them with known RefSeq annotation, then used a Bayesian bipartite network ("SAVNet") to infer association between somatic variants and tumour-specific splicing changes at these sites. They imposed strict positional constraints of only considering bases within -6 or +6 of a known or novel splice site, which allowed them to increase their power and sensitivity and so identify 14,438 SAVs, of which 13,414 were SNVs and 1,024 INDELs[284]. 49.7% of the variants they identified fell outside the canonical donor and acceptor GT and AG dinucleotides, with the next most common positions being the bases +3 and +5 relative to known or novel donor sites. A previous study by the same group using minigene reporter assays had already demonstrated the ability of alternative splicing to be influenced by bases outside of the canonical sites[285]. Variants causing exon-skipping were most commonly found in or near exons possessing features associated with exons spliced out via the exon-recognition, not intron-recognition, spliceosome mechanism, i.e. shorter exon length, higher GC content, and longer flanking intron lengths. Impairment of intron recognition was slightly more commonly observed through disruption of the 5' donor site than the 3' acceptor. Novel splice donor sites were created widely throughout known exons and introns, whereas novel acceptor sites tended to be generated from within the polypyrimidine tract of known introns. 63.3% of the genes frequently affected by SAVs (≥10 samples) were tumour suppressors. Contrary to previous similar analyses by Jung *et al.*, Shiraishi *et al.* observed disruption of tumour suppressor genes more commonly through exon skipping and alternative splice-site generation than though intron retention, though a significant proportion of events

(12%) were still due to the latter. Consistent with inactivating tumour suppressors via the triggering of NMD, transcripts with intron retention events had significantly lower expression when assayed by RNA-seq compared to WT transcripts, though exon skipping and alternative splice site usage events did also show reduced expression if their splicing changes corresponded to a frameshift. SAVs were also found in a number of oncogenes, among them a recurrent event in the exon 14 donor site of the hepatocyte growth factor receptor, *MET*, which resulted in an in-frame exon-skipping event which has been demonstrated to constitutively activate c-Met in non-small cell lung cancer[284,286].

These analyses highlight the ability of somatic variations in splicing to aid and drive cancer progression.

1.5.4 Non-coding germline variants influencing splicing can predispose to cancer

Early efforts to assess the mechanisms of non-coding mutations predisposing to cancer centred around regulatory control of gene expression. For example, through use of ChIP-seq, nucleosome occupancy and chromosome-conformation fluorescence imaging, Schödel *et al*. characterised a non-coding locus on 11q13.3 predisposing to renal cell carcinoma as being an enhancer of *CCND1* expression containing a hypoxia-inducible factor (HIF) binding motif[287]. Similarly, an intronic variant within the *LMO1* gene has been demonstrated to predispose to neuroblastoma by producing a novel *cis*-acting enhancer for the gene through creation of a GATA binding site.[288]

However, increases in the read length and depth of RNA-sequencing have led to increasing scope for non-coding variants predisposing to cancer to be able to be linked to splicing alterations. Stacey *et al.* identified 4 new loci linked to basal cell carcinoma via a GWAS involving 4,572 cases, one of which they ascertained caused intron retention from exon 8 to 10 of *CASP8.* According to RNA-seq, carriers of the risk allele had significantly reduced expression of the full-length *CASP8* transcript and concomitantly increased expression of intron 8 retention according to a multivariate linear regression (β=-0.65, P=1.7x10^{-12})[289]. They replicated their findings in RNA-seq from the GTEx project collected from sun-exposed skin samples, with intron retention and reduced expression of the canonical transcript again significantly correlating with the C risk allele (β=-0.64, P=7.2x10^{-9}). CASP8 is

a key signalling mediator in the apoptotic pathway, and variants in this region have also been linked to other cancers, including breast and melanoma[289]. The intron retention event likely corresponds to reduction in CASP8 expression via a nonsense mediated decay signal in proximity to exon 10, and its loss could render incipient cancer cells less likely to be proactively culled.

Tanha *et al.* demonstrated that homozygous or heterozygous carriers of the germline SNP rs2274407 have a worse 3 year disease free survival of paediatric acute lymphoblastic leukaemia (ALL) from a cohort of 145 cases ($P=1.9x10^{-4}$, OR=13.17, 95% CI=2.55-68.11)[290]. The variant lies in the 3' acceptor site of exon 8 of the *ABCC4* gene which encodes MRP4 ("multidrug resistance-associated protein 4"), known to be involved in the efflux of multiple xenobiotic organic anions. The SNP causes loss of approximately 300bp of exon 8 through weakening of the canonical acceptor site, and may lead to poorer clinical outcome due to reduced tolerance to chemotherapy, meaning that drug regimens might be attenuated or curtailed early for carriers of the variant before achieving a therapeutic benefit.

In relation to colorectal cancer specifically, Soukarieh *et al.* used minigene reporter assays to profile the effects of 22 germline mutations in exon 10 of *MLH1* associated with Lynch syndrome, and found that 17 of them affected splicing either through interruption of canonical splice sites, ESEs or ESSs[291]. Building on this work, Rhine *et al*. surveilled a further 36 known pathogenic exonic variants across 5 other exons of *MLH1*, and which fell outside of classical splicing-associated sequences (they defined these as falling outside of -5 to +6bp in relation to the 5' SS and -20 to +3 around the 3' SS). They found that 11 out of 36 (30.5%) of these variants significantly affected splicing, which all fell within just 2 of the exons tested (6/6 in exon 8 and 5/7 in exon 15), and that they had significantly greater predicted effects on the strength of exonic regulatory elements compared to WT sequences than exonic mutations not found to influence splicing (P=0.0280, Mann-Whitney)[292]. This highlights that the number of variants affecting splicing may be underesimated, and that synonymous and even nonsynonymous coding variants outwith classically defined splice sites may have a previously unappreciated influence on alternative splicing regulation. The authors had assumed that the variants would act by hindering intial recognition of splicing-related sequences and spliceosome assembly, however from an *in vitro* assay of the complex they discovered that 63% of variants functioned by impeding transition from complex A to complex B, and 37%

prevented activation of complex B to B*.

1.5.5 Potential therapeutic targeting of aberrant splicing in cancers

Given the role that aberrant splicing plays in cancer, multiple approaches have been taken or proposed to therapeutically target the spliceosome and its associated processes.

Cancers driven by overexpression of the c-Myc transcription factor have a hallmark of excessive transcriptional activity, with a 1.5-fold increase in absolute intracellular RNA levelsl[293]. This provides a therapeutic window whereby normal cells are able to tolerate some degree of spliceosome inhibition, whilst Myc-activated cancer cells are operating at near maximum splicing capacity. Then treated with spliceosome inhibitors, cancer cells undergo genome-wide intron retention, producing many mis-spliced transcripts and resulting non-functional polypeptides that must be removed by the cell. If the intracellular lysosomes are not able to cope, the cells will perish under conditions termed “proteotoxic stress”[294]. shRNA knockdown of the spliceosomal component BUD31 in breast cancer cell lines driven by c-Myc significantly impaired proliferation through inhibiting BUD31’s interactions with core spliceosomal components of the U2 Complex[295]. The addition of HER2 or EGFR driver constructs to the cell lines did not increase the magnitude of the response, implying that the effect is indeed c-Myc-specific. Pladienolide and spliceostatin A, small molecule inhibitors of SF3B1 derived from microbial sources, cause increased apoptosis in breast and lung cancer cell lines[296]. The SF3B1 small molecule inhibitor Sudemycin D6 has been shown to impair spliceosome function and cause synthetic-lethality in c-Myc driven cancer cell lines and mammalian models[295].

Khales *et al.* performed analysis of genome and RNA sequencing of 32 tumour types across 8,705 TCGA tumour-normal pairs and found that there was an increase in diversity of alternative splicing events of up to 30% in tumour vs normal tissue, with an average of 930 “neojunctions” of exon-exon links per tumour not observed in any GTEx normal tissue samples[297]. The authors predicted mis-splicing events which would lead to the generation of tumour-neoantigens with MHC-1 binding potential and so the ability to elicit an immune response, and found an average of 1.7 such neojunctions per tumour sample (and 0.6 SNVs per sample

predicted to cause neoantigens). Given that a number of the neoantigens were recurrent across >100 cancer samples, the authors predict that such analyses could be used to generate broadly applicable chimeric antigen receptor T cell (CAR-T) or cancer vaccine therapeutics, or they could simply be used to stratify patients by those most likely to respond to immunotherapies such as programmed cell death checkpoint inhibitors.

1.6 Hypothesis and aims of thesis

Colorectal cancer has consistently been demonstrated to have a significant heritable component, and early diagnosis and intervention can markedly improve prognosis and survival. Common sQTL SNPs with low effect size have been demonstrated to significantly contribute to other complex traits including autoimmune and neurological diseases. Aberrant somatic splicing caused by both *trans*-acting mutations in core spliceosomal components and *cis*-acting synonymous and non-synonymous variants can contribute to the progression of a number of cancers, including CRC. Certain high-penetrance germline mutations causing mis-regulation or mis-splicing of genes have been demonstrated to contribute to predisposition to a rare number of cases of certain cancers, including CRC.

Therefore this project hypothesises that common germline sQTLs may explain the mechanism of action of a significant portion of the non-coding variants predisposing to colorectal cancer. This project aims to comprehensively identify sQTLs in colonic mucosa samples of a Scottish cohort, the precise tissue of origin of CRC, and to characterise the loci involved.

The first results chapter justifies the choice of RNA quantification platform, and includes quality assessments of the data. It also contains an exploration of gender differences in gene co-expression in colonic mucosa.

The second results chapter details the results of sQTL identification via two separate and complementary algorithms. It describes the significance and effect size of the sQTLs, shows their distribution in relation to their target feature, and explores the classes of alternative splicing change that are observed. Thresholds are also applied to filter for only the highest confidence and greatest effect-size sQTLs to further characterise.

The third results chapter assesses the predicted functional impacts of variants linked to sQTLs, and uses genome-wide annotations of regulatory features such as DNase hypersensitivity and ChIP-sep of histone modifications to characterise the genomic regions which sQTLs occupy. The genomic inflation of sQTL SNPs is assessed in relation to p-values from a meta-GWAS of CRC predisposition, and examples are given of sQTLs for genes associated with CRC predisposition and progression.

Chapter 2 Methods and Data Collection

This chapter details the collection of normal colonic mucosa samples from human donors, and their RNA sequencing in two separate batches. There are further technical methods contained within each of the subsequent results chapters relating to the specific analyses carried out within them.

2.1 Donor cohorts

As this project is primarily concerned with predisposition to CRC mediated by relatively common, moderate-to-low effect size alleles, no individuals with Mendelian disorders strongly predisposing to CRC, e.g. FAP or Lynch Syndrome, were included in this study.

Where normal colonic mucosa was obtained, it was stripped from the underlying stromal and muscle layers so that subsequent RNA-seq best represented the transcriptome of the specific tissue-of-origin of CRC[35].

2.1.1 SOCCS and COGS cohort (batch 2013152)

The first batch of normal mucosa samples came from individuals routinely recruited to two on-going cohorts designed to study genetic predisposition to CRC. These are the Scottish Study of Colorectal Cancer (SOCCS) and the Colorectal Cancer Genetic Susceptibility (COGS) study, as described in previous work by Zgaga *et al.*[298] and Theodoratou *et al.*[299]. The majority of normal mucosa samples obtained from these cohorts were extracted during surgery to resect CRC tumours, which could have been of the right or left side of the colon. Approximately 3mm^2 samples were extracted from the same side of the colon as the tumour, but at least 30cm away from the tumour site to avoid sample contamination with cancer cells. Whole blood samples for genotyping were taken pre or post operatively.

Ethical and regulatory approval for COGS and SOCCS was obtained from the South East Scotland Research Ethics Committee (under project references 11/SS/0109 and 01/0/05 respectively) and the NHS Lothian Research and Development Department (with references 2013/0014 and 2003/W/GEN/05 respectively). All participants were provided with detailed information describing the study protocol and its aims, and gave informed written consent. Confidential medical and genetic information was handled in compliance with UK legislation and tissue samples were managed in compliance with the Tissue Act Scotland, 2006.

RNA from samples from 96 individuals in the SOCCS and COGS cohorts, extracted by Dr Li Yin Ooi as part of her PhD thesis, were sequenced together under the identifier of “batch 2013152”[47].

2.1.2 SCOVIDS cohort (batch 10525)

The second batch of normal mucosa samples were sourced from the Scottish Vitamin D Study (SCOVIDS). This study was designed by Dr Peter Vaughan-Shaw to investigate relationships between circulating vitamin D levels and gene expression in normal colonic mucosa of Scottish individuals. These were mainly non CRC-sufferers who were sampled in the outpatient clinic or day surgery wards. There were two “arms” to this cohort: one observational in which a single mucosa sample was taken simultaneously with a single time point of whole blood, and a second interventional arm whereby patients taking vitamin D supplements donated normal mucosa and blood samples at multiple timepoints including 0, 6 and 12 weeks.

Ethical and regulatory approval to prospectively recruit and sample participants for SCOVIDS was granted by the South East Scotland Research Ethics Committee under project reference 13/SS/0248 and the NHS Lothian Research and Development Department under reference 2014/0058.

The SCOVIDS cohort contributed 125 unique individuals to this project, which were sequenced under the identifier of “batch 10525”. Including the extra samples from multiple timepoints, there were 187 primary tissue samples sequenced in batch 10525.

In addition, there were also 18 cell line samples sequenced in batch 10525. These were from three different colorectal cancer lines (HCT116, SW480 and LS174T), each either untreated or treated with vitamin D supplementation in triplicate. These samples are not of primary concern to this thesis, however they provided a useful QC of RNA extraction and sequencing, as detailed in the first results chapter.

Between the two batches, there was a total of 301 primary and cell line RNA-seq samples, of which there were 221 unique primary donors (96 from batch 2013152 and 125 from batch 10525) on which sQTL discovery was performed. For the 50 individuals from the SCOVIDS cohort with multiple timepoints, only mucosa RNA-

seq from timepoint 0 was used for sQTL discovery so that results were not confounded by interventional vitamin D supplementation.

2.2 RNA sequencing

RNA was extracted from colonic mucosa samples using the “Ribopure RNA extraction kit” (Applied Biosystems) according to the manufacturer’s protocol. RNA integrity and yield was quantified using the 2100 Bioanalyzer®. Extracted RNA was reverse transcribed using Moloney Murine Leukemia Virus reverse transcriptase (Promega) and random primers (Promega) at 37^0C for 30 minutes and 95^0C for 5 minutes[1].

RNA samples were submitted to the Edinburgh Genomics sequencing facility, where QC, ribosomal-depletion, strand-aware library preparation and Illumina adapter ligation was performed. Ribosomal RNA was depleted using the New England Biolabs NEBNext rRNA Depletion Kit according to the manufacturer’s protocol. Samples from batch 2013152 were sequenced on the Illumina HiSeq 2500 in “rapid mode”, producing 100bp paired-end reads. Samples from batch 10525 were sequenced with 150bp paired-end reads. The mean number of reads for batch 2013152 was 130 million, compared to 155 million for batch 10525 (*Figure 2.1*).

[1]RNA was extracted for batch 2013152 by Dr Li Yin Ooi and batch 10525 by Dr Peter Vaughan-Shaw.

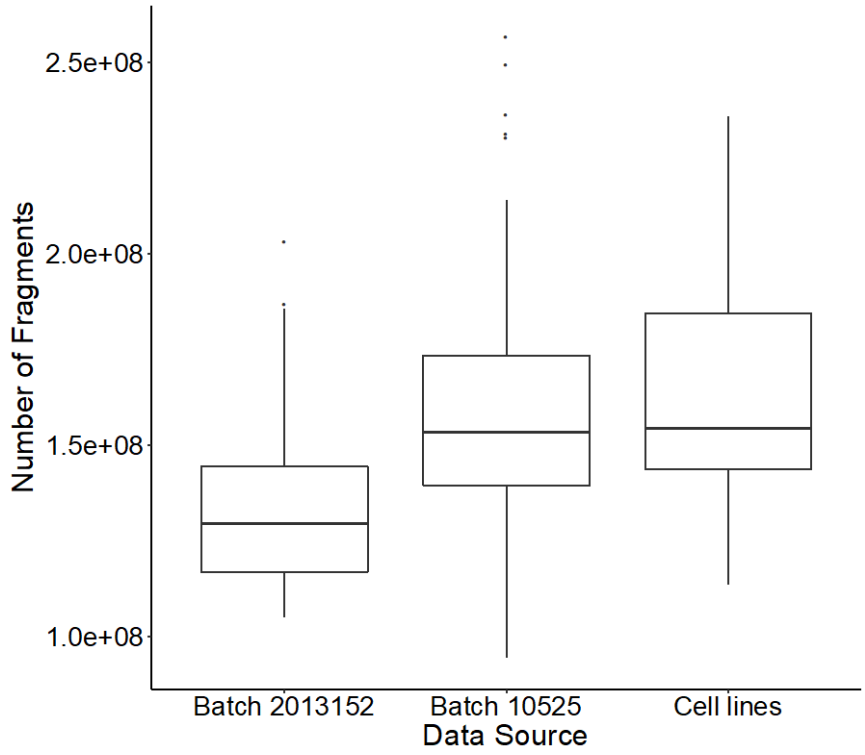


Figure 2.1 Number of fragments sequenced from each of the batches of primary samples and cell lines

2.3 Distribution of clinical metadata

Baseline clinical metadata for batch 2013152 was extracted from SOCCS and COGS databases by Dr Li Yin Ooi, with any missing data manually collected from the electronic hospital database or clinical case notes. Metadata for batch 10525 was collected during recruitment and sampling of the SCOVIDS cohort by Dr Peter Vaughan-Shaw.

Given that the majority of participants from batch 2013152 were undergoing surgery for CRC, whilst participants from batch 10525 were mainly non-diseased volunteers to the SCOVIDS study, there was a greater proportion of individuals over the age of 60 in batch 2013152 compared to batch 10525 (*Figure 2.2*).

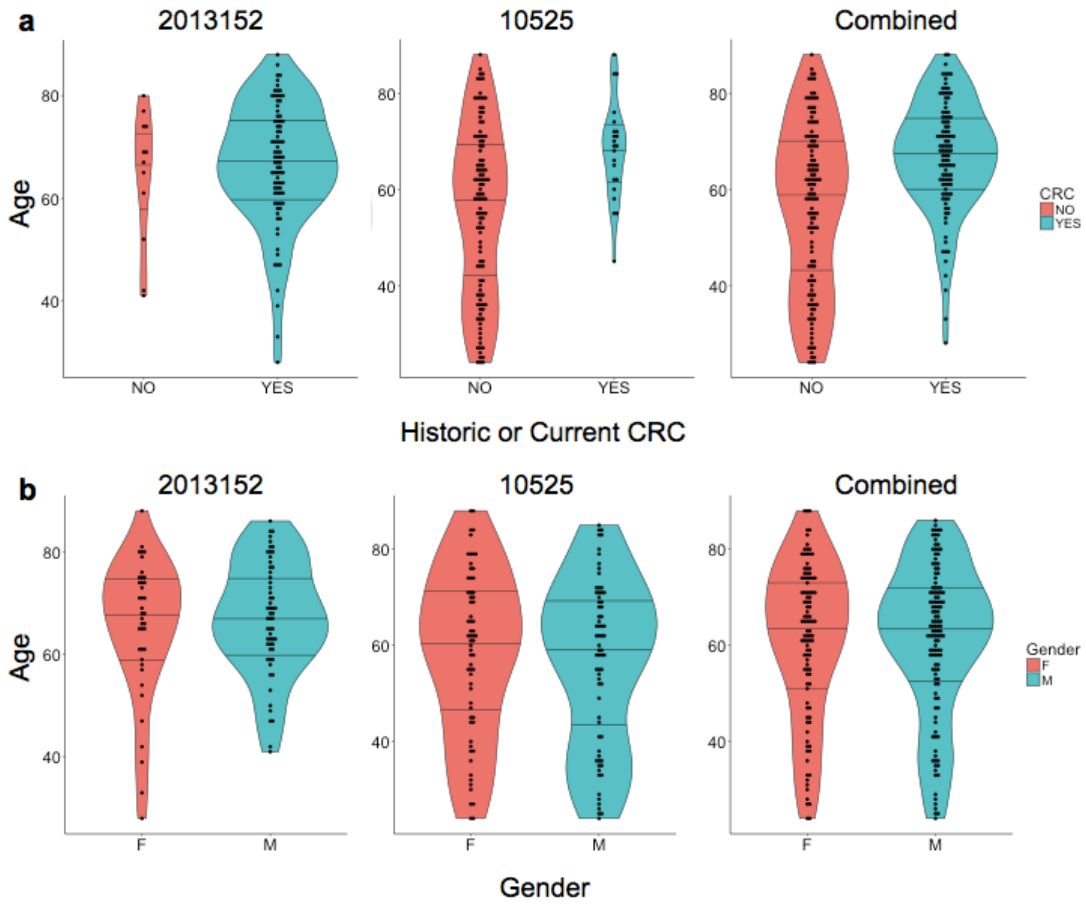


Figure 2.2 Distributions of clinical metadata
a) Distribution of ages by historic or current CRC status **b)** Distribution of ages by gender.

Right-sided cancers are less readily detected, and therefore are less likely to be operated on[48]. This is reflected in the distribution of sides within batch 2013152, with 34 right-sided samples compared to 62 left-sided (*Figure 2.3*). Given that all of the SCOVIDS participants were sampled as outpatients as opposed to open or laparoscopic surgery, all of their samples were obtained from the left side of the colon via rectal sampling. This resulted in an overall skew towards left-sided samples within the whole of the dataset used by this project. Whilst Dr Li Yin Ooi detected some differences in expression between left and right-sided samples by microarray, only 12 genes had a log fold change of greater than 2.0[47]. Given this small difference, samples from both sides of the colon were analysed together in this study to increase power for the identification of sQTLs.

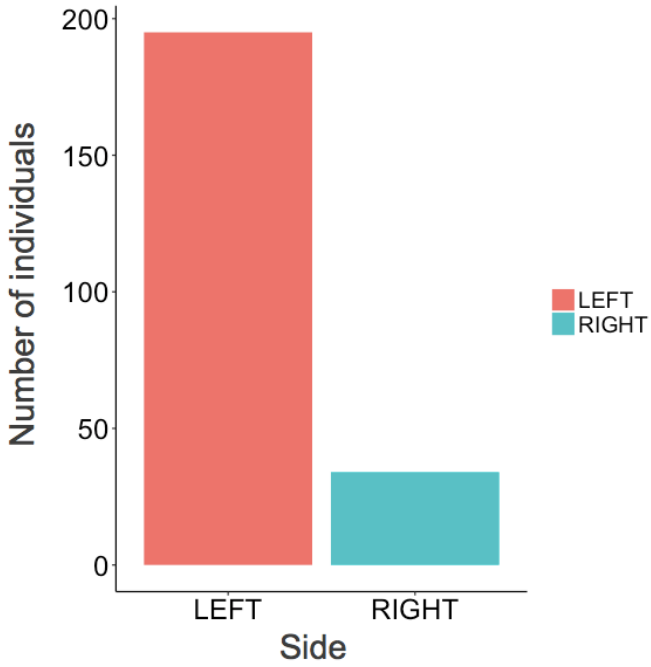


Figure 2.3 Number of individuals sampled from each side of the colon.

Chapter 3 Processing of Expression Data

3.1 Introduction

This chapter deals primarily with QC analysis of the RNA-seq data, as any conclusions drawn in this thesis rest upon the accurate quantification of transcript-level expression.

During the course of this project, there was an advance in the field of expression analysis as "alignment-independent" RNA-seq quantification tools, such as Kallisto[300] and Salmon[301], were released and began to be widely adopted. These methods markedly changed the way that read counts were assigned to genomic features, with the intention of producing more accurate transcript level quantification. Therefore the decision was taken to apply these tools to the expression data from both batches of sequencing and compare their output to the previous alignment-based quantification, which used STAR[302] to align reads to the genome and Cufflinks[303] to quantify feature expression.

The chapter then explores the batch effects resulting from two separate sets of samples being sequenced approximately 2 years apart. Two different methods, ComBat and PEER, are applied in an attempt to remove batch effects, however it is found that their outputs are incompatible with the sQTL detection tools used downstream.

A network analysis is also carried out to check for any systematic differences in gene co-expression between male and female samples, so that their inclusion together in the same sQTL discovery analysis can be justified.

3.1.1 Quantification of mRNA expression

RNA-seq consists of quantitative sequencing-by-synthesis from a cDNA library generated from the mRNA molecules present in a cell. Cells are first lysed, and RNA is isolated and fragmented. Random primers are added upon which DNA polymerase acts to produce single-stranded complementary DNA (cDNA)[304]. The RNA template is then digested and another round of polymerisation produces double-stranded DNA fragments[305]. The desired fraction of fragment lengths are selected, usually by gel electrophoresis. In the Illumina workflow, specific "Y-

shaped" adapters are then ligated to the ends of the fragments. The library is washed over a flow cell and the adaptors anneal to complementary probes immobilised to the cell surface. The adaptors also act as primers, and "bridge-amplification" creates dense local clusters of the target sequence to increase the signal for imaging[306]. Fluorescently labelled nucleotides are washed over the flow cell, which are blocked at their 3' end meaning the sequence can only be extended by a single base at a time. The process of adding and imaging single bases is continued for a number of cycles until the desired read length is achieved. Paired-end sequencing operates on both ends of each fragment, and usually leaves a tract of unsequenced bases in between. The two reads are termed "mate pairs", and there is an "insert size distribution" across the library relating to the size of the intervening sequence, which depends on the size-selection step of the protocol and the length of the feature (transcript) being sequenced. Paired-end sequencing aids accuracy when mapping reads back to a genome or transcriptome as it places additional positional constraint on where the sequenced ends must align relative to each other, and also allows for more efficient detection of intron-exon boundaries when their relative position is significantly different from that expected given the insert size distribution.

Despite random priming being used, there are still sequence-specific biases which can be introduced into RNA-seq, depending on inherent inequalities between the likelihood for particular motifs to have different affinity kinetics for their own complementary primer, and so be more or less likely to be sequenced[307]. There are obvious length and expression biases in that longer and more highly expressed features have greater likelihood to be sequenced - though the relationship between these parameters and the probability of being sequenced is not precisely linear, and are also dependent on the total library size[304].

After sequencing, there are two different approaches for quantifying RNA-seq reads against features such as genes or transcripts. Traditionally, a two-step process has been employed whereby reads are firstly aligned to the whole genome using a tool such as STAR[302] or TopHat[308] which produce BAM alignment files detailing the most likely genomic locations of each read, to which a second algorithm such as featureCounts[309] or Cufflinks[303] is applied to assign the alignments to the genomic features they overlap i.e. genes or transcripts.

Two algorithms were developed in 2016, Kallisto and Salmon, which take a different approach of performing the quantification in a single step against a reference transcriptome. These “alignment-independent” approachs, termed “pseudo-alignment” by Kallisto and “quasi-mapping” by Salmon, are both faster and more storage efficient because large BAM files containing information about each individual read do not need to be produced - the output is simply a list of features contained in the reference transcriptome and their corresponding quantification values. Utilising only a single process also means there is less opportunity for technical biases occurring from using two separate alignment and quantification algorithms.

Kallisto generates k-mers from sequenced reads and quantifies them against a de Bruijn graph of all k-mers in the transcriptome, combined with paths detailing the edges between k-mers representing known transcripts[300]. It builds equivalence-classes of k-mers which correspond to the same transcript, and prunes absent k-mers from the reference transcriptome to increase the speed of inference.

Salmon also considers k-mers of reads, and quantifies them against a hash-table index of k-mers from the reference transcriptome in a two-phase process[310]. The first phase streams reads in a random order and uses stochastic Bayesian inference to construct various bias-models and build estimates for the abundance of each equivalence class of observed sequences[301]. In the second phase it iteratively updates these estimates via expectation maximisation using the bias models until the abundance estimates converge. Unlike Kallisto, Salmon does track the position and orientation of reads within the transcripts it assigns them to, and so is able to construct more fine-grained bias models for the beginning, middle and end of transcript sequences, which Bohnert *et al.* showed can be differentially affected by technical biases with successfully sequenced fragments being more likely to originate from the start or end boundaries of transcripts[311].

The unit of quantification for RNA-seq can simply be “counts”, X_i, i.e. the number of reads assigned to a given gene or transcript feature, i. Because alignment-independent algorithms quantify transcripts probabilistically, they are able to assign non-integer counts for a feature because a read does not have to be strictly assigned to a single transcript, and instead can have a non-zero probability of having originated from multiple different transcripts.

Alternatively, units such as FPKM or TPM can be used to normalise expression of features within each sample for the read depth and feature length, given that longer features have greater likelihood to produce sequenced reads. The "effective length" ($\tilde{l}_i$) of a feature is the number of base positions from which a read of the observed length could potentially have been produced if it did originate from that feature. It is defined as the feature length minus the mean of the fragment length distribution for that sample plus 1, because there must be at least 1 base for the random priming to have begun from. For features with length less than the mean fragment length, the effective length as calculated by this method is 1.

$$\tilde{l}_i = l_i - \mu_{FLD} + 1$$

The feature-specific biases which Salmon calculates are incorporated into the estimates of effective transcript-length for each sequence in the reference transcriptome, which subsequently affects the likelihood of assigning a read to a particular feature.

FPKM (fragments per kilobase of sequence per million mapped reads) is the observed number of counts for a feature corrected for the length of the feature in kilobases, then corrected by the total library size (N reads), which is finally scaled by a factor of 1 million to make the values more interpretable[312].

$$FPKM_i = \frac{X_i}{(\frac{\tilde{l}_i}{10^3})} . \frac{1}{N} . 10^6$$

If FPKM were to be a true measure of "relative molar RNA concentration", then for a given transcript species in an analysis its average should be invariant, however Wagner ***et al.*** demonstrated that this is not always the case[313]. TPM (transcripts per million mapped reads) is a different measure which has become widely adopted because it avoids such statistical discrepancies. TPM is calculated by correcting the counts observed for a feature by its effective length, then dividing by the sum of all corrected feature counts in the library, and scaling by 1 million for interpretability[314].

$$TPM_i = \frac{\frac{X_i}{\tilde{l}_i}}{\sum_j \frac{X_j}{\tilde{l}_j}} . 10^6$$

As a result, TPM is effectively a measure of the expected number of reads that would be attributable to transcript i, given the abundances of all other transcripts in the sample, for each 1 million fragments sequenced from the library.

The majority of RNA in a cell is ribosomal (80-90%[315]), whereas the desired target of RNA-seq is mature mRNA molecules. Therefore mRNA is enriched either by affinity capture of the poly-adenylated tails of mature transcripts with thymine oligos[316], or by treating cell lysates with DNA oligos with affinity for ribosomal sequences then introducing RNase H enzymes which digest DNA-RNA duplexes, thus depleting ribosomal RNA[317,318].

3.1.2 Genome Builds

The initial round of RNA-seq analysis and sQTL identification performed on batch 2013152 was aligned against the GRCh37 reference genome. The transition from GRCh37 to GRCh38 incorporated 1,158 different fixes, including the filling of 198 gaps and 34 tracks of missing sequence, and has been demonstrated to lead to improved mapping rates to exomes which benefits more accurate RNA-seq analysis[319]. Therefore once RNA-seq was obtained for samples from batch 10525, both batches were aligned and quantified against the newer GRCh38 release - a decision which had also been taken by other large consortia such as the 1000 Genomes Project[320].

3.1.3 Network analysis

Weighted gene correlation network analysis (WGCNA) is a tool for inferring the presence of coordinated programmes of gene expression in cells and tissues[321]. The repeated correlation of a set of genes across many samples implies that their expression is under shared control via a particular regulatory system. Given the established differences in incidence of colorectal cancer by gender[1,9], WGCNA was performed to test for any differences in gene expression control active between males and females in the colonic mucosa. WGCNA was used in order to complement differential expression analyses which have previously been carried out. The thesis of Dr Li Yin Ooi found 23 genes more highly expressed in males and 22 more highly expressed in females when using the limma package[322] to analyse microarray expression quantification from colonic mucosa samples from 64 males and 51 females. However, of those 45 differentially expressed genes, only 6 were not from the sex chromosomes (*OSCP1, DPM3, CSTF3, ZMYND12, NLRP2, COPS8*), implying that mainly global gender-specific differences, rather than colon-specific, may have been observed by differential expression[47]. WGCNA will allow the testing of whether any modules of genes are under control of different regulatory pathways in the colonic mucosa of males and females.

3.2 Methods

3.2.1 Genomic alignment of reads

Both batches were aligned using a one-pass alignment to GRCh38.p10 using STAR version 2.5.1[302] using default parameters[2].

3.2.2 Quantification of RNA-seq using Cufflinks

Feature expression was quantified from the subsequent bam files against the Ensembl gene build v88 using the Cufflinks pipeline[303] with default parameters[3].

When Cufflinks quantifies features from bam alignment files, it can assign 4 different statuses to the resulting expression value:

- OK: deconvolution successful
- LOWDATA: too complex or shallowly sequenced
- HIDATA: too many fragments in locus
- FAIL: when an ill-conditioned covariance matrix or other numerical exception prevents deconvolution

Any RNA-seq expression value emanating from a transcript not assigned as "OK" by Cufflinks was excluded from the analysis, which constituted 0.0523% of the transcripts.

3.2.3 Salmon expression quantification

Alignment-free quantification of RNA-seq fastq files was performed using the quasi-mapping mode of Salmon version 0.8.0[301].

Reference Transcriptomes

Salmon requires a reference transcriptome against which to quantify expression. Both cDNA and ncRNA (non-coding RNA) transcriptomes were downloaded from the Ensembl ftp server and manually concatenated to provide a comprehensive reference (*Table 3.1*).

[2] Credit Dr Alison Meynert for STAR alignment.
[3] Credit Dr Victoria Svinti and Dr Alison Meynert for Cufflinks quantifications.

Genome Version	Gene Build	Reference Transcriptome	URL
GRCh38	88	cDNA	ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz
GRCh38	88	ncRNA	ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz

Table 3.1 URLs for Reference Transcriptomes

Salmon workflow

The Salmon workflow firstly required construction of a custom index of the reference transcriptome to facilitate its alignment-free algorithm. 59 transcripts from GRCh38.v88 shorter than the minimum k-mer length utilised by Salmon for quantification (31nt) were flagged by the package, meaning that reads were unlikely to be quantified against those transcripts. The default k-mer length for quantification could be changed, but the default setting was used for comparability with other studies.

Salmon was run in "quasi-mapping" mode, quantifying feature expression from fastq files directly against the combined cDNA and ncRNA reference transcriptome. Salmon was run with library option "ISR" to denote that read pairs were facing inwards towards each other, were stranded, and with read1 in the reverse orientation.

All three of Salmon's in-built bias detection and correction algorithms were employed, relating to sequence bias, GC content, and positional biases. The bias models are trained on the first 1 million reads streamed by Salmon, and because the reads are supplied in a random order these are assumed to provide a representative sample of all fragments present in the library. A variable-length Markov Model (VLMM, similar to that originally developed by Roberts *et al.* for the 2013 eXpress algorithm[323]) is used to detect any sequence-specific bias at the 5' and 3' ends of fragments which can occur due to random hexamer priming causing fragments beginning or ending with certain motifs to be preferentially sequenced. The second correction is applied to account for the likelihood of a sequence being observed given its internal GC content, given it has been demonstrated that correcting for this bias at the level of individual fragments improves quantification accuracy for samples with appreciable levels of GC bias, without impairing accuracy if no

significant bias is present[301,324]. Sequence bias and GC content bias are related, but not identical, phenomena, therefore Salmon learns three different fragment-GC bias models relating to the start, centre and end of the sequences of fragments. The final correction attempts to take account of coverage biases arising from a non-uniform distribution of fragment start-sites for a particular feature, learning different models for different lengths of transcripts.

3.2.4 Analysis of Salmon quantification success rate

Salmon reports the unique identifiers of any reads which were not able to be quantified against the reference transcriptome. It was desired to test what proportion of these effectively "unmapped" reads were able to be aligned genome-wide by STAR, and where in the genome they fell. Any fragments where neither mate pair read were quantified by Salmon were extracted from bam files of STAR genomic alignments using the "*FilterSamReads*" function from Picard tools version 1.139[325]. The percentage of Salmon unmapped reads that were able to be aligned by STAR was calculated as the number of unique Salmon unmapped reads present in STAR-generated bam alignments per sample divided by the total number of Salmon unmapped reads per sample.

This percentage was then further partitioned by whether STAR aligned the reads within or outwith exonic sequences, which were defined as the union of all exon coordinates according to a biomaRt[326,327] query of Ensembl gene build 88[328] mapped to GRCh38.p10[329]. The "*CountReads*" function from GATK tools version 4.0.0.0[330] was used to tabulate the numbers of Salmon unmapped reads which fell within or outwith the exonic regions. These numbers could be potentially misleading for two reasons. Firstly, reads which aligned across the boundaries of exonic regions were counted twice by the CountReads algorithm; once as exonic and once as intergenic. In order to solve this problem, any reads which mapped across the boundaries were removed such that the percentages were only calculated based only on reads which unambiguously mapped within or outwith exonic regions. The second issue arises in that the bam files containing unmapped Salmon reads had some secondary mappings, where the same read could be mapped multiple times. This means that if there were different rates of secondary mappings within exonic regions than outside them, then the percentages for each region could potentially be skewed. However the resulting metrics are still useful for observing an overall trend.

The "*idxstats*" function from SAMtools version 1.6[331] was used to count the number of reads which were genomically aligned to each sequence contig (chromosome or alternative scaffold) by STAR.

When quantifying the number of reads mapped by STAR to ribosomal RNA genes, rRNA sequences were defined from biomaRt[326,327] query of Ensembl gene build 88[328] mapped to GRCh38.p10[329] for the exons of any gene with the biotype "rRNA". 5kbp windows were added to either side of these sequences, in order to capture the repetitive regions that often surround the annotated locations of rRNA genes, and therefore are likely to also attract repetitive ribosomal reads[332]. The "*CountReads*" function from GATK tools[330] was again used to record the numbers of reads falling within or outwith these rRNA regions.

For visualisation in the Integrative Genomics Viewer (IGV)[333], bam files were converted to bigwig read-density tracks using the function "*bam2bw*" from the package cgpBigWig[334].

3.2.5 Principal Components Analysis

Principal components analysis (PCA) was run separately at the gene-level and transcript-level on log and quantile normalised Salmon counts. Logarithms were taken of each count with the formula:

$$X' = log_2(X + 1)$$

such that values less that one would remain positive after log transformation. Quantile normalisation was performed using the "*normalize.quantiles*" function from the preprocessCore R package version 1.32.0 with default settings[335]. The "prcomp" function[336,337] from base R[338] was used to calculate principal components, with the arguments "*centre=TRUE*" and "*scale=FALSE*".

3.2.6 Batch correction with ComBat and PEER factor residuals

ComBat

ComBat batch correction was run on log_2 and quantile normalised transcript-level TPM values as quantified by Salmon. 1.46% of the transcripts were discounted for exhibiting variance of 0.0. ComBat was run from the sva (surrogate variable

analysis) package version 3.18.10[339–341] with the argument "*par.prior=TRUE*", indicating that parametric adjustments as opposed to non-parametric would be used.

PEER factor correction

Analysts from the Genotype-Tissue Expression Consortium (GTEx) report a relationship between the number of samples in an analysis and the number of unknown factors which should be corrected for using PEER (probabilistic estimation of expression residuals) in order to maximise the number of cis-eQTL associations identified. For a sample size n<150 they suggest using 15 factors, for 150≤ n <250 to use 30 factors, for 250≤ n <350 to use 45 factors and for n≥350 to use 60 factors[342]. Therefore 30 factors were used when running PEER on the 221 samples in this analysis.

Firstly an arbitrary low-expression threshold was set requiring transcripts to have ≥6 counts in ≥10% of the samples. This retained 156,806 out of 217,082 transcripts (72.2%). Then TMM (Trimmed mean of M-values[343]) normalisation was performed using the "*calcNormFactors*" function from the edgeR package version 3.16.5[344,345]. This was followed by using the "*rntransform*" function from the GenABEL package version 1.8.0[346] to produce a normal distribution of expression values by matching the inverse normalised rank of each transcript to the corresponding quantile of a normal distribution. 30 unknown factors were estimated using the "*PEER_update*" function from the peer package version 1.0[347,348]. As a comparison, PEER factor correction was also performed on non-normalised data with only the transcript expression threshold imposed.

3.2.7 Differential network analysis using WGCNA

WCGNA was performed on the male and female samples from batch 10525 only, to ensure there would be no confounding effects from differences in library sizes between batches which could affect correlations. The larger of the two batches was chosen as it contained more individuals and a roughly equal split between the two genders. The network analysis was carried out on 124 individuals after a single outlier identifiable when PCA was performed on batch 10525 was removed (sample MD14398_A, female aged 31, left-sided sample, no history of CRC, BMI 22.4). Salmon gene-level counts were filtered to select for the 5,000 most highly expressed

genes and the 5,000 genes with the greatest variance to reduce the likelihood of false-positive correlations between lowly expressed features or irrelevant correlations between stably expressed genes. The intersect of these two sets was taken, which resulted in 4,400 genes. Their Salmon counts were $\log_2$ and quantile normalised as described previously before correlations were calculated to increase comparability between samples prepared from different libraries.

Firstly, a male-specific correlation network was generated from the 66 male samples. Pearson correlation coefficients were calculated pairwise between all genes, and a network was constructed whereby the genes formed nodes and edges between nodes were weighted according to the correlation coefficients.

Construction of an effective weighted gene correlation network is reliant on achieving an approximately “scale free” topology[349]. Networks approximating this classification have the characteristic that the distribution of clustering coefficients follows a power law, whereby sparsely-connected nodes are common and there are increasingly fewer, more highly-connected hub nodes which are aggregated into “modules”[350]. This model approximates the regulatory hierarchy predicted to be active in most biological systems[351]. In order to achieve scale free properties, the correlation coefficients between genes are raised to a power. Given that the vast majority of correlations will be <1.0, raising such numbers to a power will considerably reduce their magnitude and serve to selectively trim poorly supported edges, by penalising them disproportionately more severely (e.g. 0.99^8 = 0.923, whereas 0.25^8 = 0.0000153). Each time a network is constructed, the appropriate power is selected by estimating the fit of the resultant correlation matrix to a scale-free topology, with the authors of WGCNA recommending that the coefficient of the fit should achieve approximately 0.9[349]. The networks are unsigned such that any significant correlation between genes causes them to be co-located into the same module of highly correlated genes (as opposed to signed networks, whereby only positively correlated genes are clustered)[349].

In order to detect differences between the expression networks of males and females, a second consensus network was then constructed using a combination of both genders’ expression data, and the resulting modules of highly connected genes was compared to the first male-specific network. Any modules which are discordant

between the individual and consensus gender networks are inferred to contain genes which are under different regulatory regimens in males and females.

3.3 Results

3.3.1 Correlation between Cufflinks FPKM and Salmon TPM

Median Cufflinks FPKM per transcript and median Salmon TPM per transcript were compared for the 96 samples from batch 2013152 (*Figure 3.1*), producing a correlation coefficient of Rho=0.678, P=<2.2e^{-16}. The rank-based Spearman correlation coefficient was used, because although the TPM and FPKM quantification scales are not directly comparable, the rank of features between each methodology should be similar if the two methods perform comparably. When displayed as a scatter plot on a log scale, it is clear that the best-fit line deviates from y=x because of a skew of transcript quantifications which are higher according to Salmon than Cufflinks.

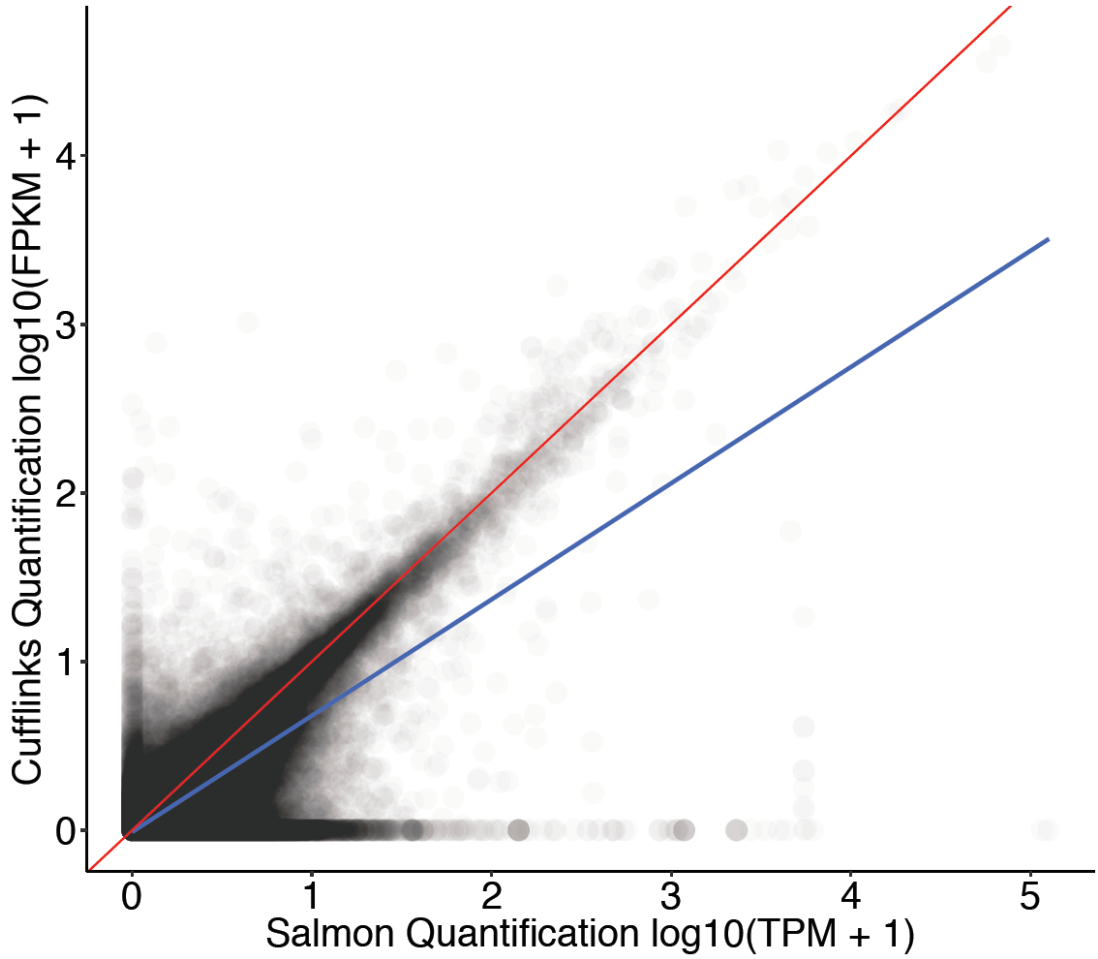


Figure 3.1 Cufflinks against Salmon transcript-level quantifications.
Best fit line (lm method from ggplot2) in blue, line of y=x in red.

3.3.2 Quantification success rates of Salmon and STAR

There was a greater success rate of Salmon quantification for reads from the 18 cell line samples which were sequenced as part of batch 10525 than either of the batches of primary samples (*Figure 3.2*). There was a greater STAR mapping success rate than Salmon for reads from primary samples from both batches. However, the STAR mapping success rate for cell lines was lower than that achieved by Salmon.

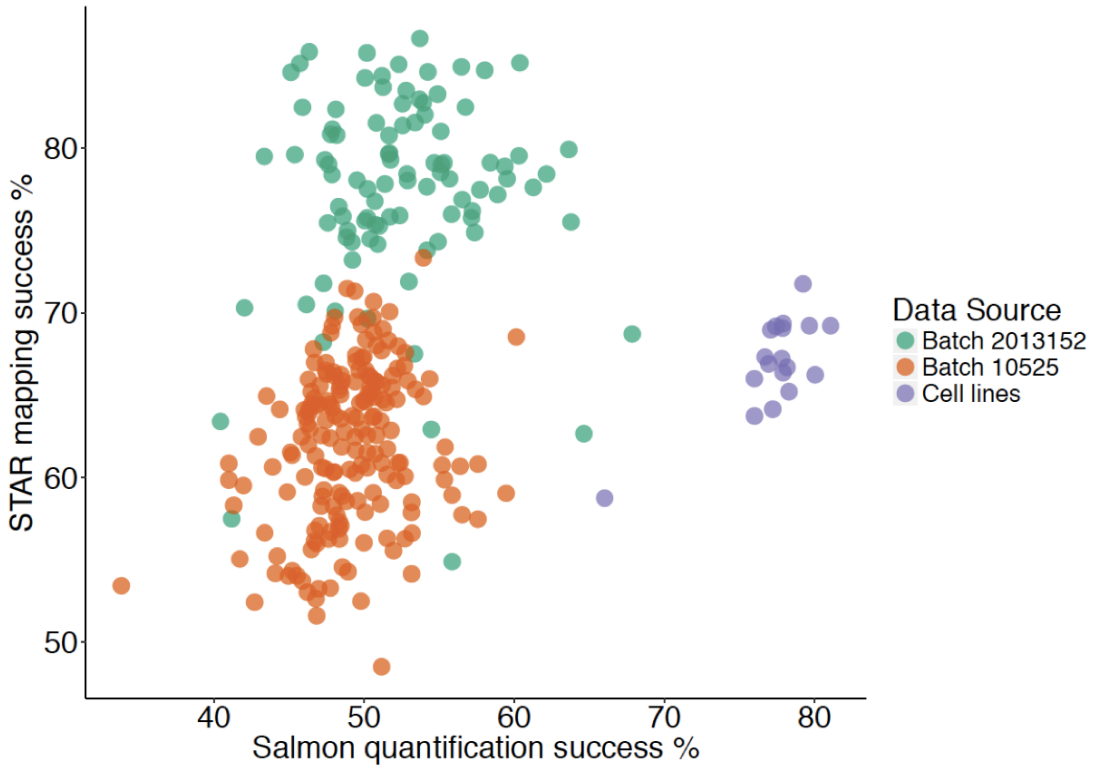


Figure 3.2 Mapping success rates of Salmon and STAR across batches.

The Salmon quantification success rates of approximately 50% of reads for primary samples were lower than expected, however its performance of near 80% from cell line samples indicates that it is not an issue of the algorithm itself. Therefore the issue of low quantification success of reads by Salmon in primary samples was further investigated.

When STAR-aligned bams were queried for the read IDs which were un-quantifiable by Salmon, it was found that a greater percentage of the Salmon unmapped reads from batch 2013152 were able to be mapped by STAR than from primary or cell line samples from batch 10525 (*Figure 3.3* a). The main cause of reads failing to be mapped by STAR was the reads being too short (*Figure 3.3* b). The majority of reads which were not quantifiable by Salmon but which were mapped by STAR were mapped to regions outside the transcriptome, as defined by a union of all exonic regions (*Figure 3.3* c). For all batches, the proportions of secondary mappings produced by STAR was greatest in the bams containing reads not quantifiable by Salmon (*Figure 3.3* d), implying that these reads are enriched for lower-quality or more repetitive sequences.

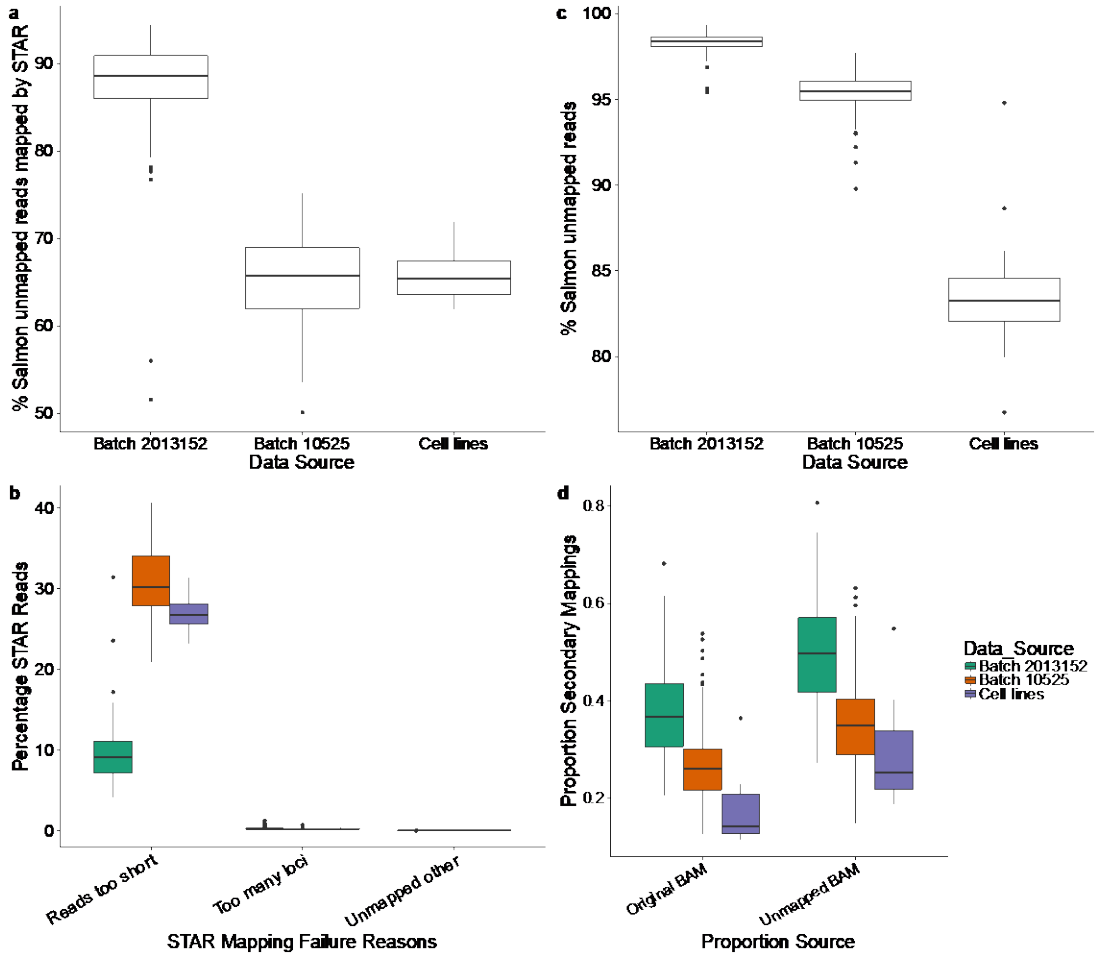


Figure 3.3 STAR mapping successes for reads not quantifiable by Salmon
(a) Percentage of reads not quantified by Salmon which were able to be mapped by STAR **(b)** Percentage of STAR reads per batch which failed for either being too short, mappable to too many potential loci, or being unmapped for other reasons **(c)** Percentage of reads not quantified by Salmon which were mapped to non-exonic sequences by STAR **(d)** Proportions of secondary mappings of reads in original STAR bam files, and bams containing only the reads not quantified by Salmon but mappable by STAR.

Figure 3.4 plots samples ordered by the percentage of unquantifiable Salmon reads which were mapped to exonic regions by STAR. The samples fall into distinct clusters, with cell lines having the largest percentage, then batch 10525 of primary samples followed by batch 2013152.

Figure 3.5 displays read counts as opposed to the percentages, and samples are ordered by the total number of reads quantifiable by Salmon. This makes the boundaries between groups of samples less defined; the cell lines are less well clustered because although they all possessed the greatest percentage of reads quantifiable by Salmon, some primary tissue samples possessed more overall reads

than certain cell line samples. It can also be observed that primary samples from batch 2013152 tended to contain fewer reads than batch 10525 which were unquantifiable by Salmon but could be aligned to exonic regions by STAR, despite having a greater percentage score according to this metric than batch 10525 as demonstrated in *Figure 3.3* c; a discrepancy resulting from batch 2013152 having a lower total read count.

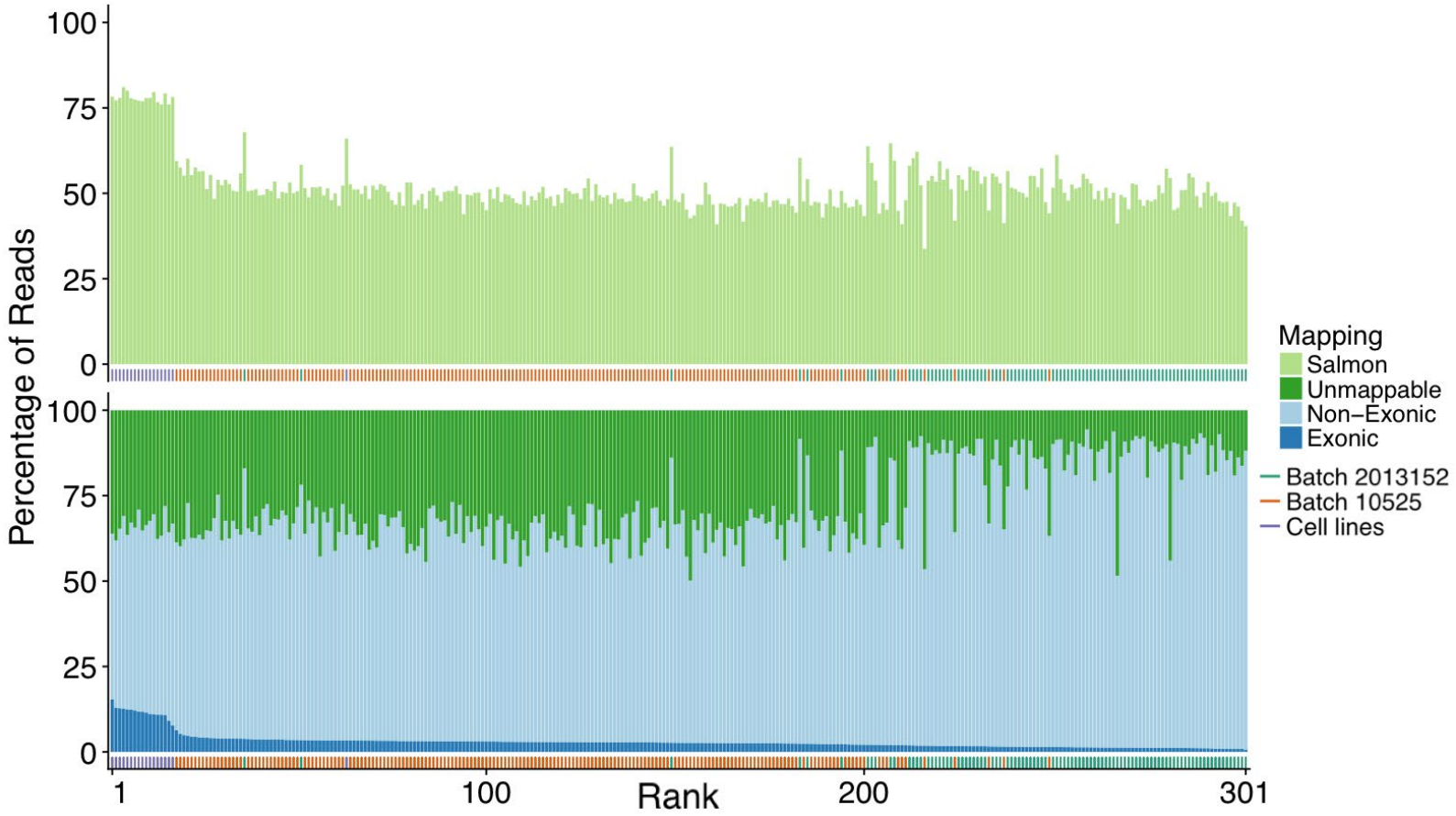


Figure 3.4 Percent mapping successes of samples ordered by unquantifiable Salmon reads mapped exonically by STAR
(Upper) Salmon quantification success in light green **(Lower)** Of the reads which were not able to be quantified by Salmon, the percentage mapped by STAR onto exonic regions are in dark blue, percentage of reads mapped to non-exonic regions in light blue, and also unmappable by STAR in dark green. **(Tick marks)** Indicate batches and cell lines.

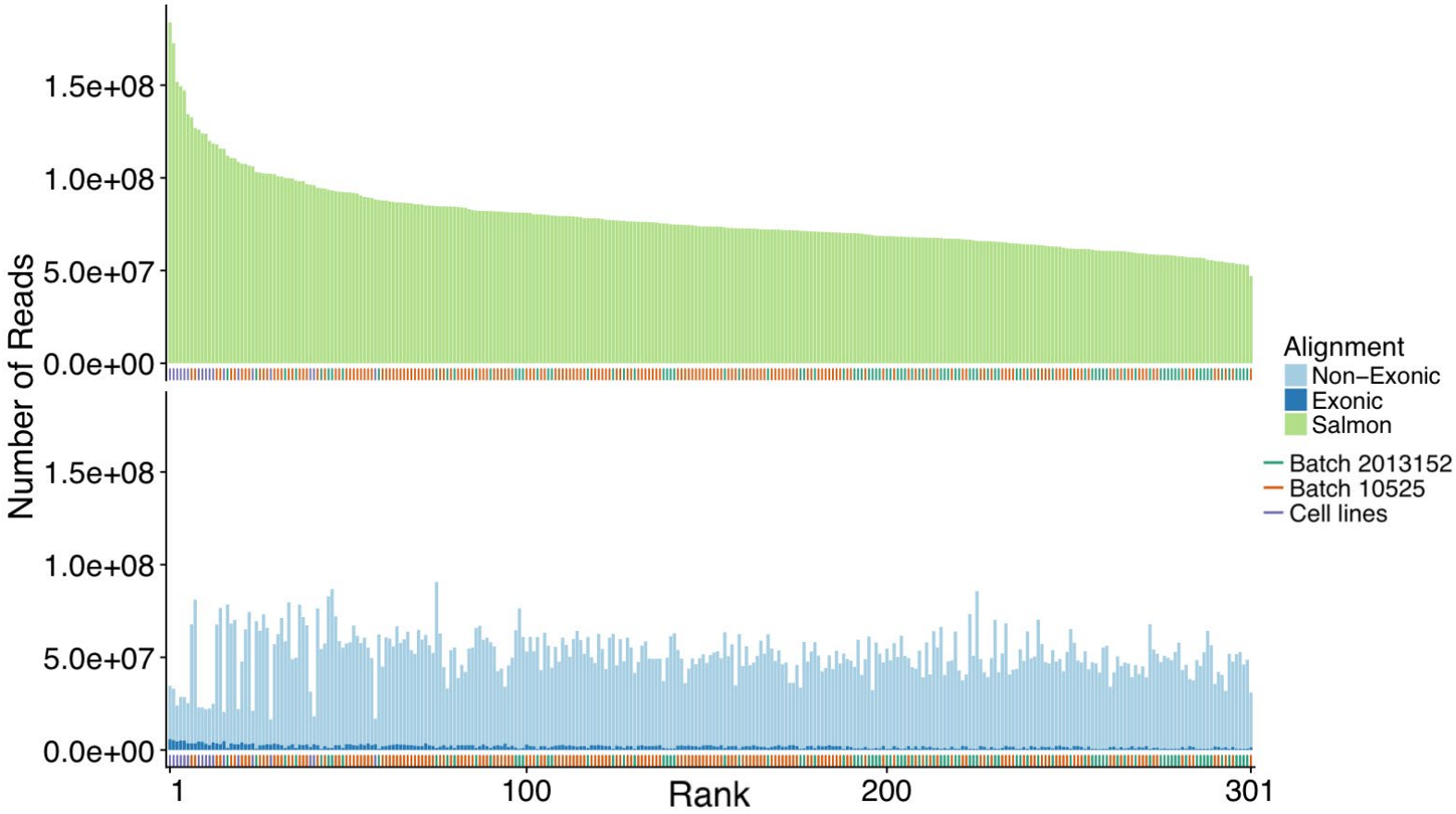


Figure 3.5 Number of reads quantified or mapped per sample ordered by total number of Salmon quantified reads
(Upper) Number of reads quantified by Salmon in light green **(Lower)** Of the reads which were not able to be quantified by Salmon, the number of reads mapped by STAR onto exonic regions are in dark blue, number of reads mapped to non-exonic regions in light blue, **(Tick marks)** Indicate batches and cell lines.

Given their lengths, there was a greater number of reads than would be expected aligned by STAR to chromosomes 14, 17, 21 (*Figure 3.6* solid line). There was also unexpectedly large numbers of reads mapping to the alternative chromosome 21 scaffolds GL000220.1 and KI270733.1 in primary samples, but not in cell lines (*Figure 3.6* solid line). These five sequences together attracted the greatest numbers of reads which were aligned to the genome by STAR but were unquantifiable by Salmon in relation to the reference transcriptome (*Figure 3.6* dashed line). A median of 48.8% of the reads from batch 2013152 samples were mapped by STAR to the ribosomal sequences on these five chromosomes +/- 5kbp windows (Figure *3.7* dashed line). The media was 34.2% of reads for batch 201525, and 14.7% for cell lines (Figure *3.7* dashed line). Large numbers of reads mapping to these regions can clearly be seen in screen shots from the Integrative Genomics Viewer (IGV)[333] of representative samples from each data source (**Figure** *3.8*).

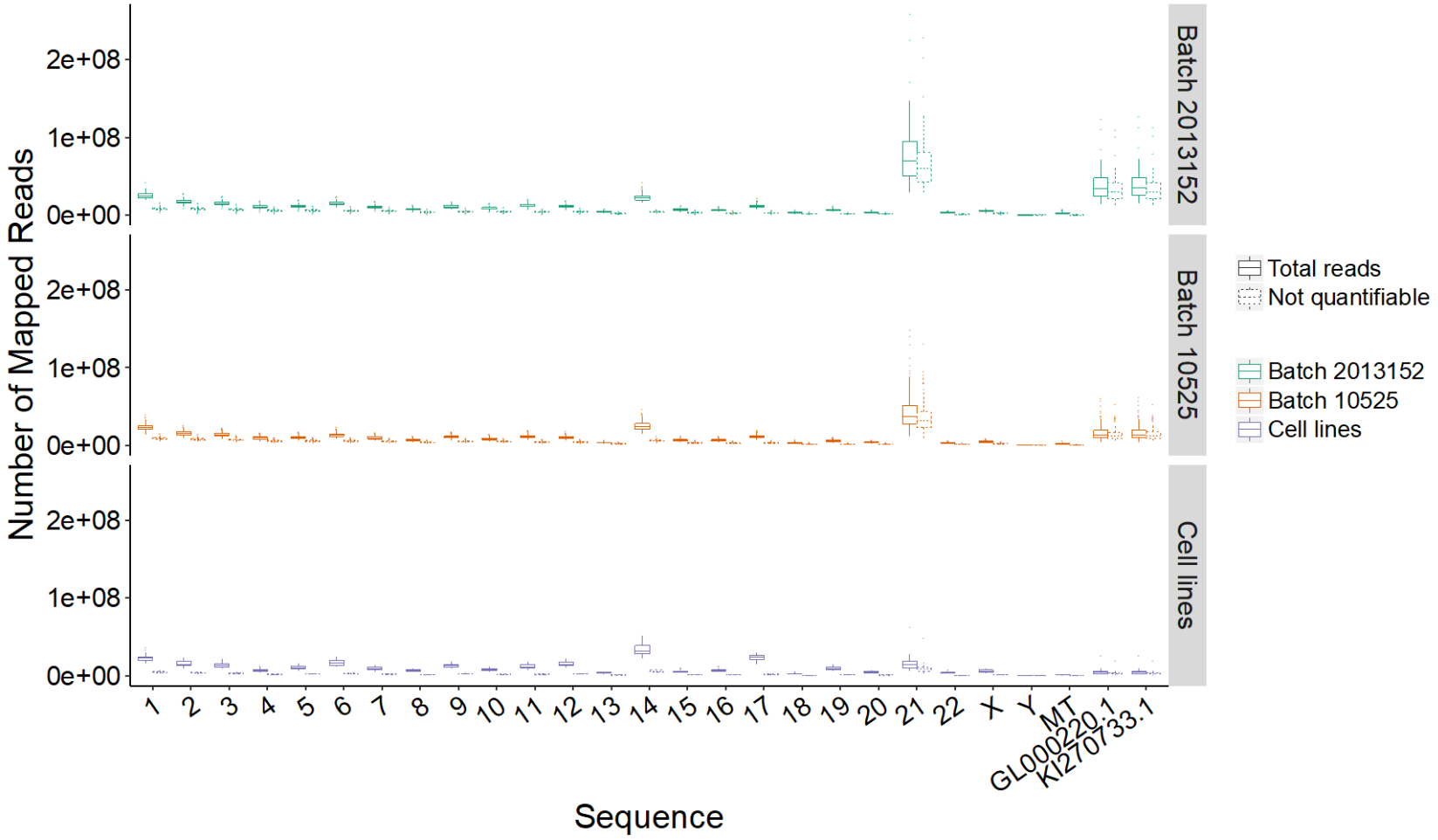


Figure 3.6 Distribution of reads STAR-aligned reads per chromosome by batch.
Solid lines indicate all reads, dashed lines indicate those reads which were not able to be quantified by Salmon.

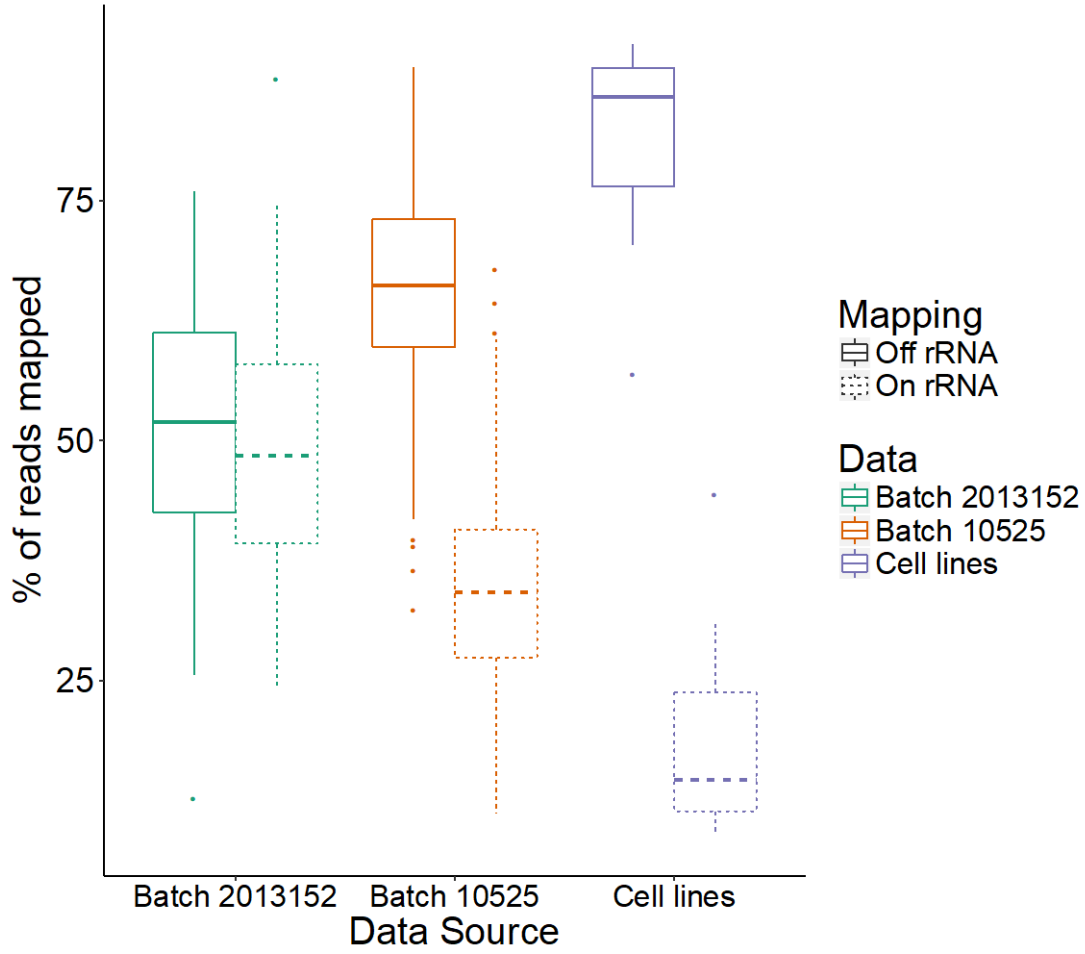


Figure 3.7 Reads mapped to genomic regions encoding ribosomal RNAs.
"rRNA" refers to regions comprising any exons of rRNA genes on chromosomes 14, 17, 21, GL000220.1 and KI270733.1 plus or minus a 5kbp window.

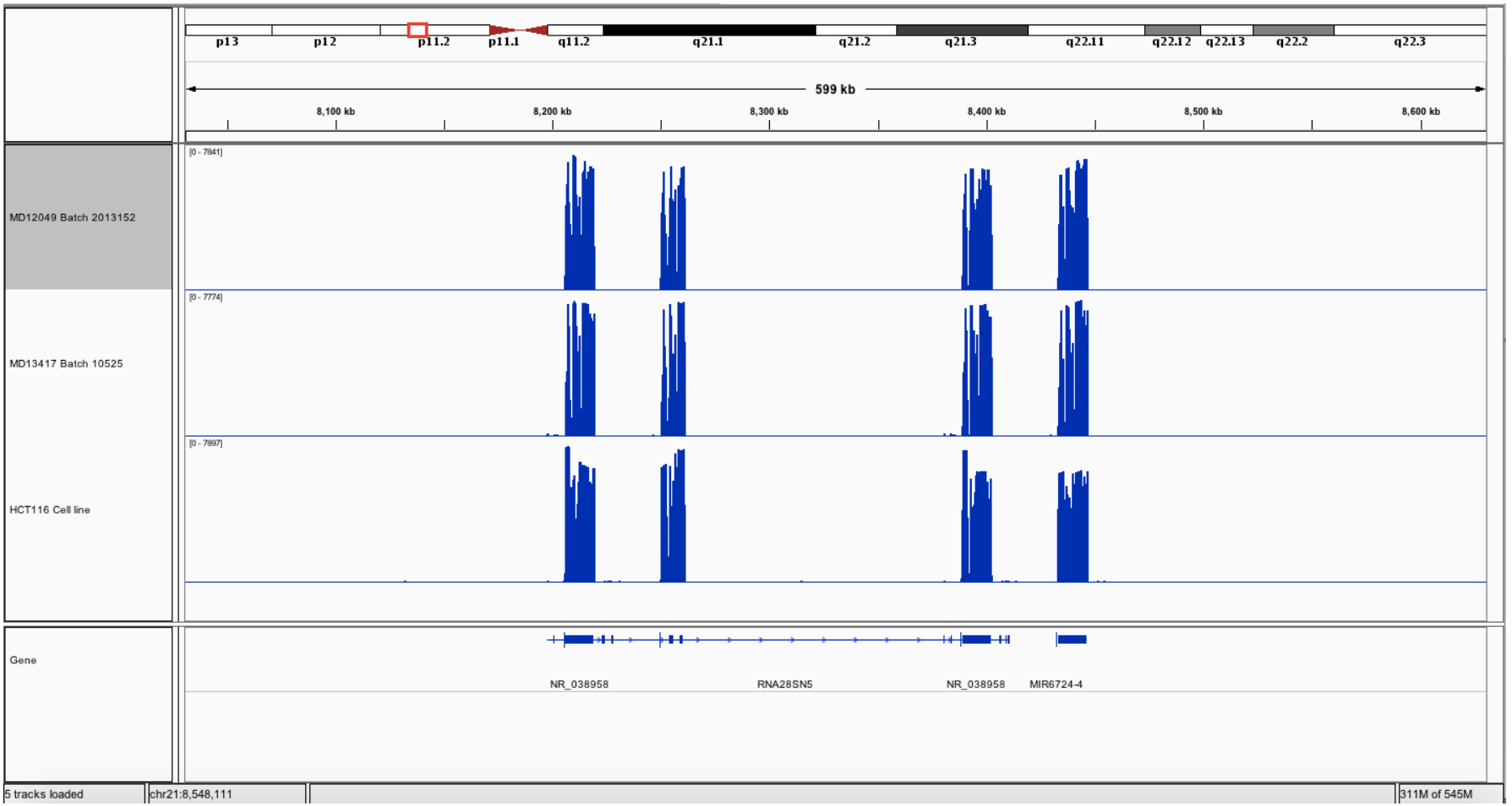


Figure 3.8 IGV screenshot of chromosome 21, coordinates 8,030,572-8,630,975.
Reads from three representative samples are shown; MD12049 from batch 2013152, MD13417 from batch 10525, and a sample from the HCT116 cell line.

3.3.3 Principal components analysis

PCA demonstrates clear differences between batches and between primary tissue samples and cell lines

20% of the variance in expression between all samples analysed can be explained by the first two principal components of Salmon transcript-level counts (*Figure 3.9*). Component 1 captures the difference between primary tissue and cell line samples, whilst component 2 captures the variance due to batch effects (*Figure 3.10*).

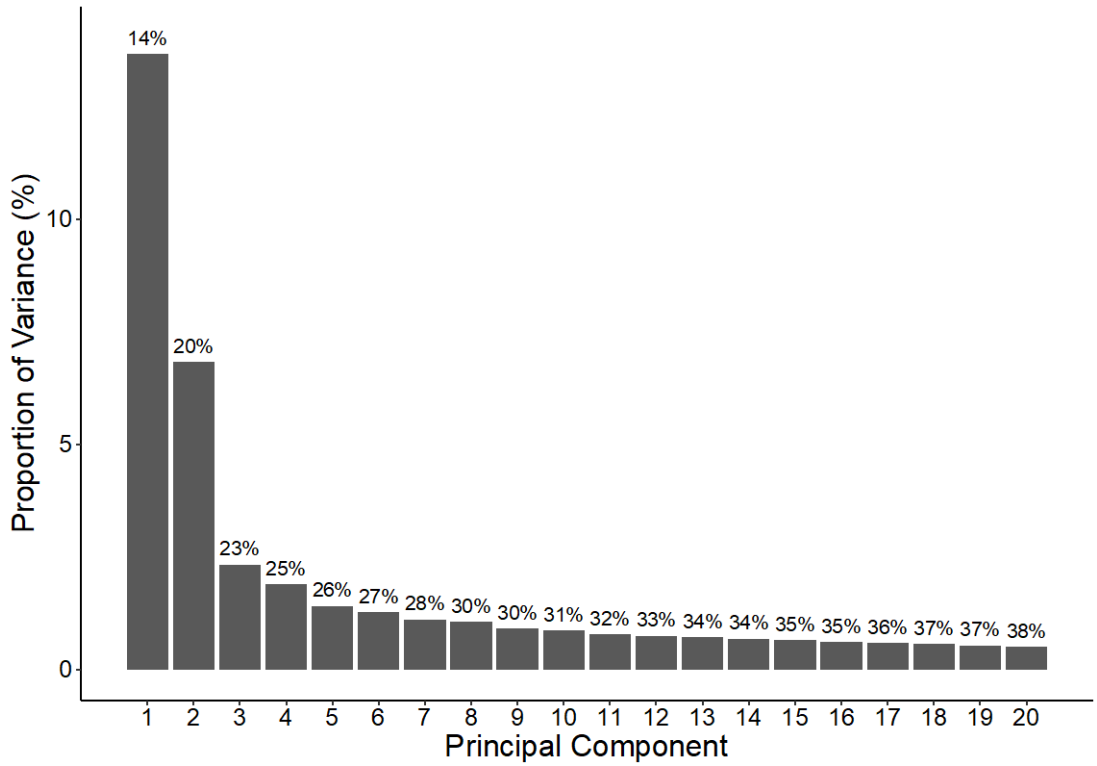


Figure 3.9 Proportion of variance explained by first 20 principal components
Derived from Salmon transcript-level counts (log2 and quantile normalised) for 283 primary and 18 cell line samples**.** Numbers above bars indicate cumulative percentage of variance explained.

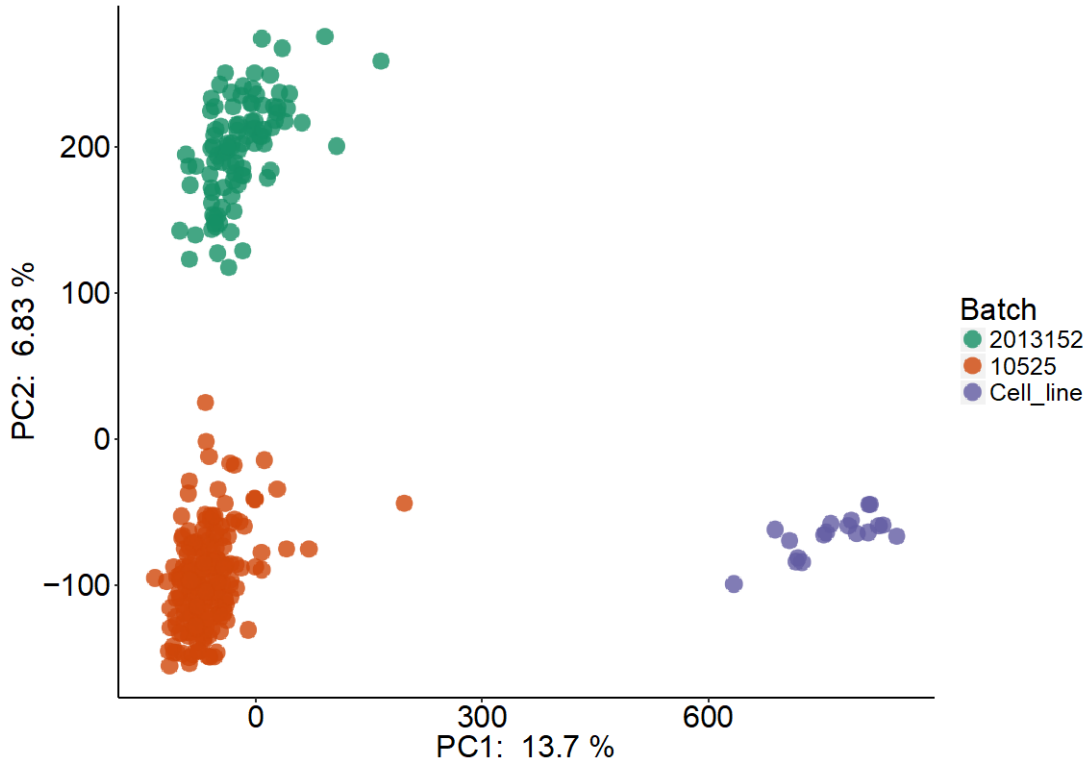


Figure 3.10 First two principal components
Derived from Salmon transcript-level counts (log2 and quantile normalised) for all 301 samples both primary and cell lines.

No obvious separation of primary samples beyond 1st Principal Component

When principal components were calculated using transcript-level counts from just the 221 primary samples which would be used for identifying sQTLs, the first component corresponding to batch-separation explains the most variance (9.26%), This was more than 4-fold greater than the second largest component (2.17%) and as much as components 2-7 combined (which cumulatively explained 9.31%). There is no clear separation in of the first 5 components apart from PC1 (*Figure 3.11*).

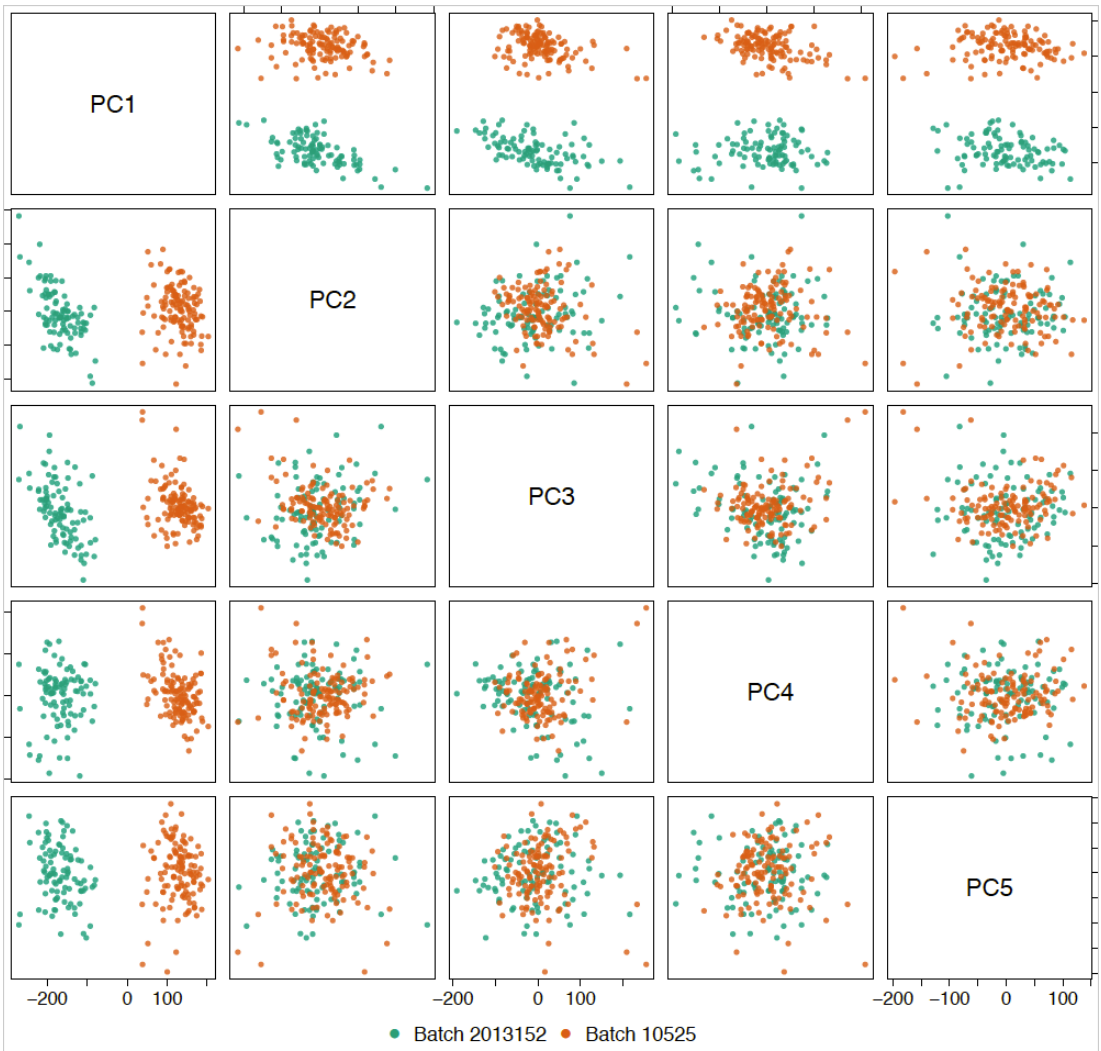


Figure 3.11 First 5 principal components derived from Salmon transcript-level counts (log2 and quantile normalised) for all 221 primary colonic mucosa tissue samples. Coloured by batch.

Gender effects are visible via gene-level PCA of a single batch, with strongest effects caused by genes with the 500 largest variances

There was expected to be a gender-separation uncovered by PCA, however it was not apparent when analysing all 221 primary samples together. Batch effects may have dominated the variance components and masked the contribution of other factors. Lowly expressed transcripts may also have produced noise capable of obscuring genuine separations. So PCA was performed again using gene-level expression from just the 125 samples from batch 10525. This identified one outlier, sample “MD14398” (*Figure 3.12*), though there was no clear indication in their metadata why they should have presented an atypical gene-level expression profile that wasn’t apparent in the transcript-level analyses of all 221 primary samples (Female aged 31, left-sided sample, no history of CRC, BMI 22.4).

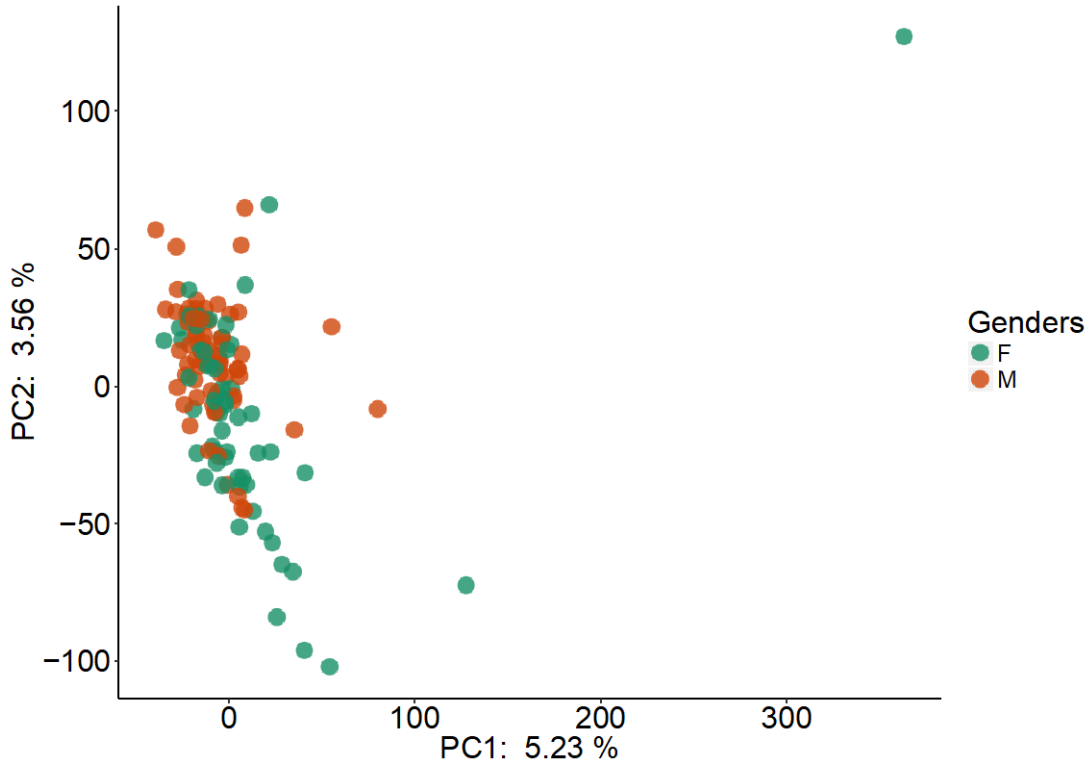


Figure 3.12 First two principal components derived from Salmon gene-level counts (log2 and quantile normalised) for 125 primary colonic mucosa tissue samples from Batch 10525.

After discounting the female sample MD14398, gene-level PCA was able to identify a gender separation in component 3. To increase the clarity of the separation further, analysis was re-performed using only the genes with the 500 greatest

variances across the 124 samples (*Figure 3.13*). This threshold was modelled on that used by the "plotMDS" function of the edgeR differential expression package[345,352]. The separation between female and male samples became sufficiently pronounced to identify one sample which had been mislabelled as female.

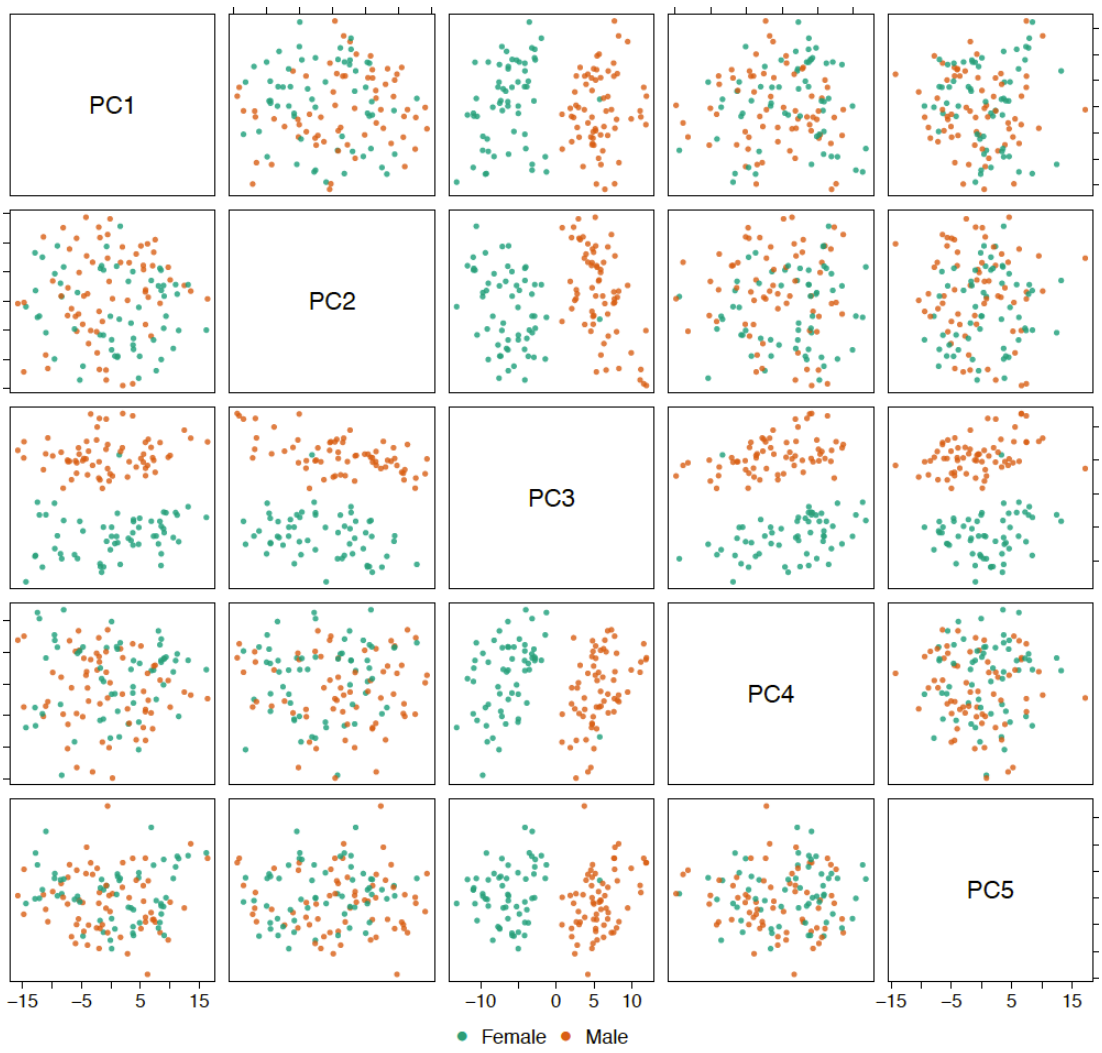


Figure 3.13 First 5 principal components derived from Salmon gene-level counts (log2 and quantile normalised) for 124 primary colonic mucosa tissue samples from Batch 10525.
Only the genes with the 500 greatest variances were used. Coloured by gender.

In order to confirm that this separation was due to the influence of gender-specific expression, the same analysis was repeated with any genes from the X or Y chromosomes which were in the list of top 500 variances removed (see *Table 3.2*). As expected, this removed any gender separation from the data (not shown).

Gene	Chromosome	Ensembl_ID
DDX3X	X	ENSG00000215301
FLNA	X	ENSG00000196924
HEPH	X	ENSG00000089472
MAOA	X	ENSG00000189221
OGT	X	ENSG00000147162
POF1B	X	ENSG00000124429
RPL10	X	ENSG00000147403
CHANGE	X	ENSG00000198034
SLC25A5	X	ENSG00000005022
TMSB4X	X	ENSG00000205542
XIST	X	ENSG00000229807
AC010970.2	Y	ENSG00000225840
TXLNGY	Y	ENSG00000131002

Table 3.2 13 genes from X and Y chromosomes which featured in the top 500 greatest variance genes across 124 samples from Batch 10525

3.3.4 Batches converge after correction with ComBat

Following batch correction with ComBat[340] there was still a single principal component which dominated the variance, accounting for 12.45%, compared with the next largest of 5.15% and 4.11% respectively. However, when the principal components are plotted, the two batches can be seen to have notably converged in PC1 compared to previous analyses, with some samples now overlapping at the peripheries of each batch (*Figure 3.14*).

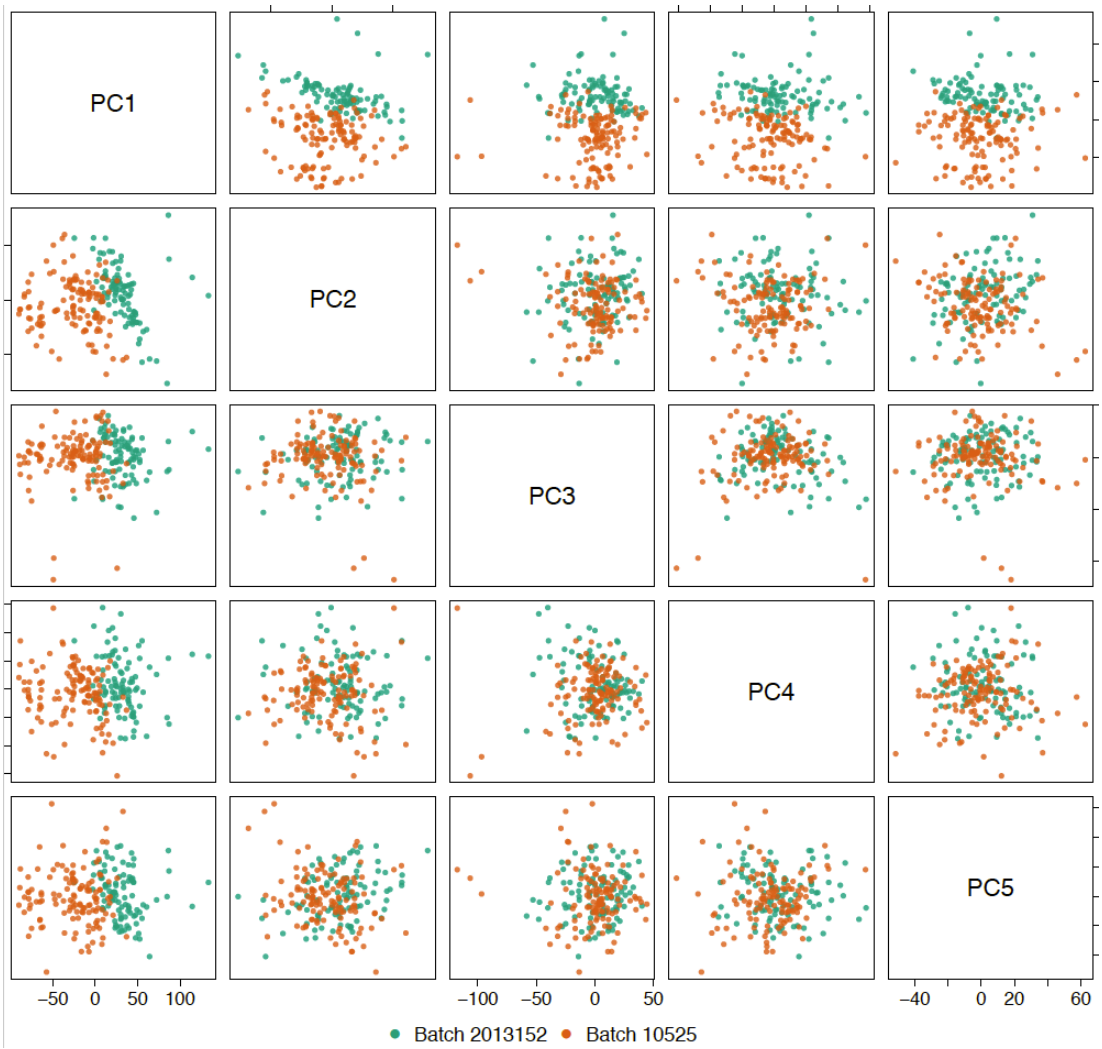


Figure 3.14 First 5 principal components derived from ComBat corrected Salmon transcript counts (log and quantile normalised).

Negative values introduced following ComBat correction made it unsuitable for use with sQTL discovery tools

After batch correction with ComBat, 14.05% of all resulting expression values were negative, and 57.0% of transcripts had at least one negative value across the 221 samples (Figure 3.15). The magnitude of the majority of the negative values was not large, however excessive data manipulation would have been required to make all the values positive and therefore compatible for input into sQTL detection algorithms.

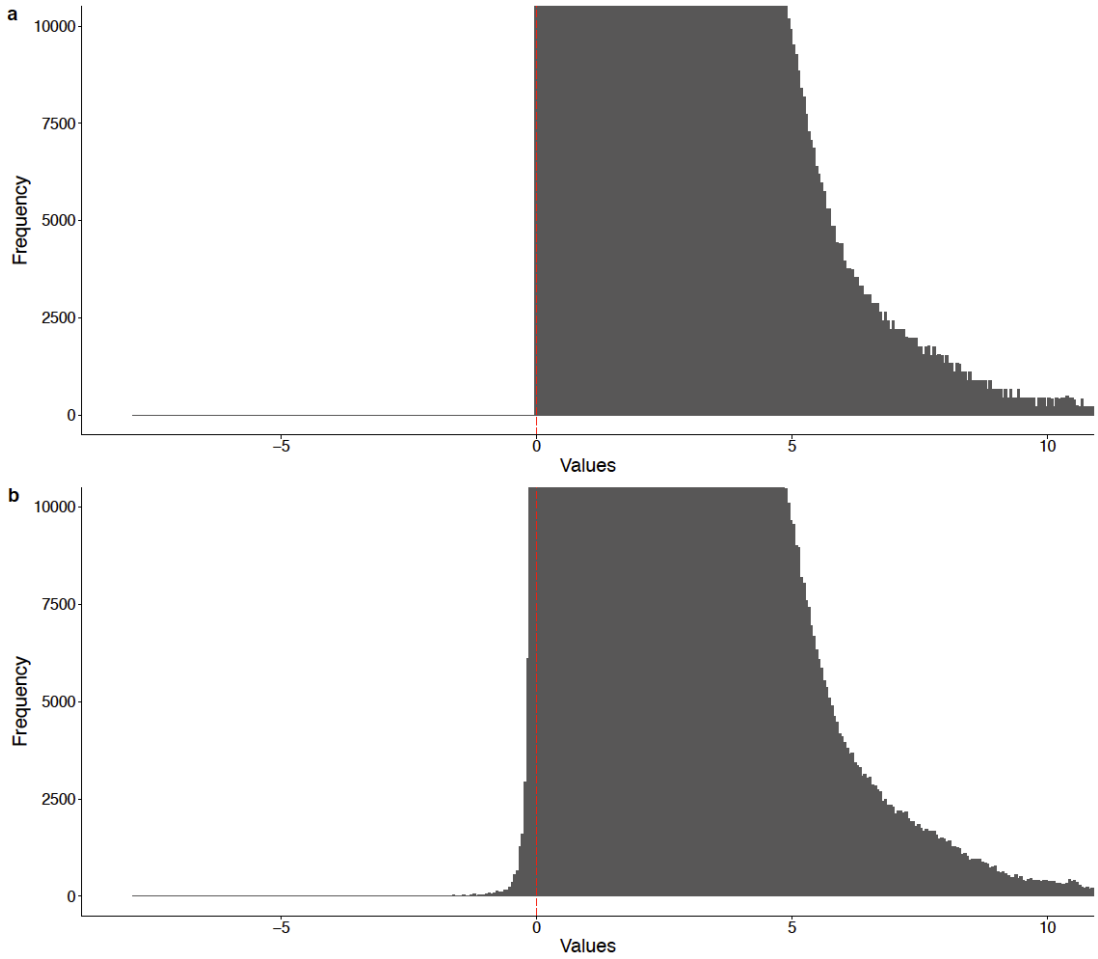


Figure 3.15 ComBat batch correction introduces negative values
a) Log and quantile normalised transcript-level counts from 221 primary samples. **b)** ComBat batch-corrected transcript-level counts. Y axes are attenuated for visibility of lower frequency bars.

3.3.5 Negative PEER factor residuals

PEER factor correction produces a list of covariates which can be used in conjunction with linear models to correct for hidden confounding variables in expression data. These could have been applicable to FastQTL, but the sQTLseekeR algorithm was not able to account for covariates at time of analysis[247]. PEER also returns residuals from the factor analysis, which can be considered a matrix of corrected values. These were plotted, however the process introduced large numbers of negative values into the data, whether PEER was run on normalised or non-normalised expression counts, therefore it was also incompatible with sQTLseekeR (*Figure 3.16*).

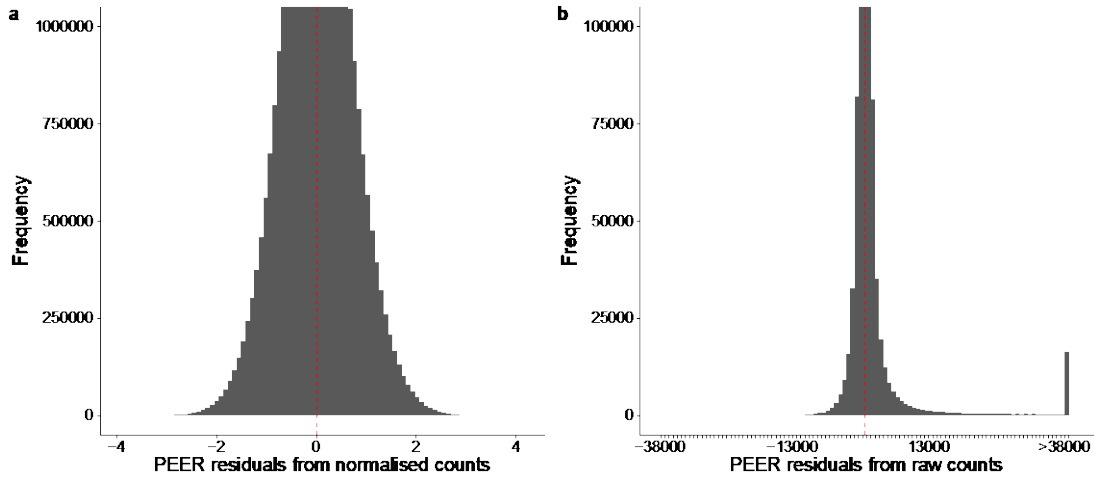


Figure 3.16 Distributions of PEER factor analysis residuals run on normalised and raw counts.
Y axes are attenuated for visibility of lower frequency bars.

3.3.6 Differential network analysis between genders

WGCNA was firstly used to generate a male-specific gene co-expression network from 66 samples from batch 10525. Agglomerative hierarchical clustering was performed via UPGMA (unweighted pair group method with arithmetic mean) using Euclidean distance between gene expression of samples as the distance matrix[353]. No clear outliers were observed in the resulting unrooted dendrogram, nor any biases in relation to age or BMI (*Figure 3.17*).

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Euclidean distance between two vectors, p and q

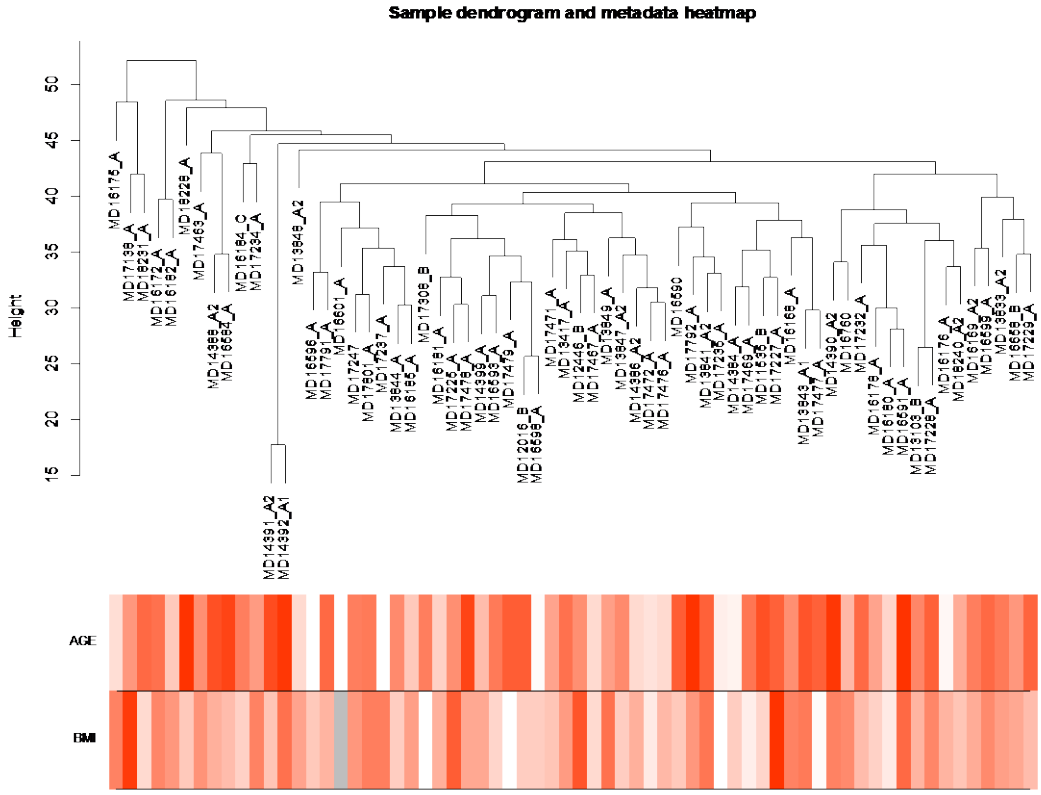


Figure 3.17 Unrooted dendrogram of 66 male samples from batch 10525.
Ages range from 24 (light red) to 86 (dark red) and BMI from 18.3 to 48.9. Grey bar indicates BMI was unavailable.

A power threshold of 8 was chosen as it achieved a fit to a scale-free topology of approximately 0.9 (*Figure 3.18*). It can bee seen that as the power which correlation coefficients were raised to increases, the mean connectivity of the network decreases as more edges are trimmed.

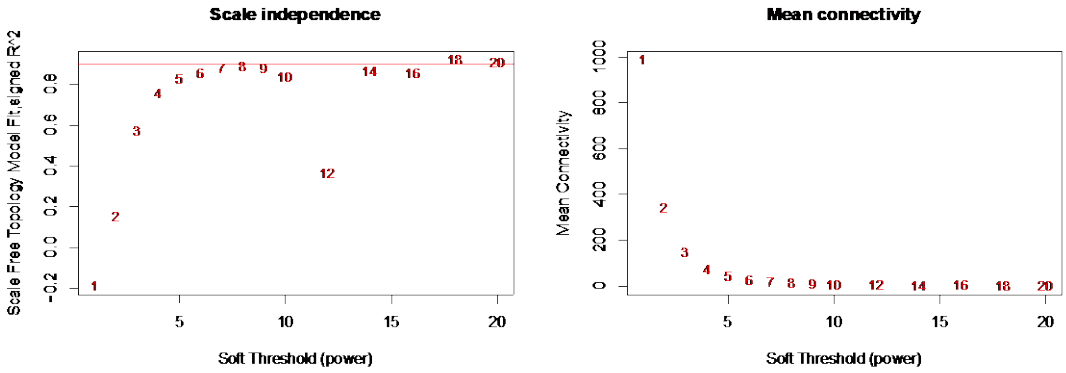


Figure 3.18 Fit to scale-free topology and mean connectivity of networks from 66 male samples with edges raised to various powers

The male-specific network was constructed using the "*blockwiseModules*" function from the WGCNA package version 1.66[349] with a minimum module-size threshold of 30 genes and default parameters for merging of related modules. Modules are constructed via hierarchical clustering of genes, where dissimilarity between two gene vectors is calculated as 1- ω_{ij} (where ω_{ij} is the Topological Overlap Matrix (TOM) as defined by Kaufman and Rousseeuw[354]) and genes are assigned to modules with high TOM similarities. 3,709 of the 4,400 genes with greatest expression and variance were able to be assigned to 18 modules, as shown in *Table 3.3*, leaving 691 unassigned.

Module	Genes	Module	Genes	Module	Genes
A	879	G	204	M	111
B	346	H	183	N	107
C	316	I	156	O	56
D	290	J	150	P	49
E	271	K	125	Q	47
F	255	L	125	R	39

Table 3.3 Numbers of genes per module for male-specific correlation network

A consensus network was then constructed from the colonic mucosa gene expression counts from 66 males and 58 females. Samples were again clustered according to Euclidean distance to check for outliers, and none were observed since the single female outlier identified via PCA of this batch had already been removed. The same power of 8 was chosen to raise the correlation coefficients to in order to achieve an approximately scale-free topology. The "*blockwiseConsensusModules*" function was used to create consensus modules using male and female expression

with ≥30 genes per module. 3,074 genes were able to be assigned to 17 modules (*Table 3.4*), leaving 1,326 unassigned.

Module	Genes	Module	Genes	Module	Genes
A	996	G	145	M	77
B	314	H	140	N	49
C	308	I	124	O	48
D	179	J	95	P	46
E	171	K	90	Q	39
F	165	L	88		

Table 3.4 Numbers of genes per module for male and female consensus correlation network

Pairwise Fisher's tests were computed to assess the comparability of modules between male-specific and consensus networks and whether the numbers of genes they shared were greater than would be expected by chance. The majority of modules from the male network had one primary module from the consensus network to which their gene content significantly corresponded (*Figure 3.19*). The male-specific module "Q" with 47 genes had no module in the consensus network that it significantly corresponded to. It shared 20 genes with the largest 996 gene module "A" from the consensus network, and shared 25 with the list of genes unassigned in the consensus network. The lack of correspondence could be because this module contained relatively few genes, however other modules of similar size did have consensus modules to which they corresponded very closely: male module "H" shared 37 of 39 genes with consensus module "Q"; male "P" shared 48 of 49 genes with consensus "N"; and male "O" shared 46 of 56 genes with consensus "O". Therefore it can be deduced that the male-specific "Q" module ceased to be detectable as a single entity in the consensus network. This is likely the module that was lost when 18 became 17, and the genes previously contained within it ended up being categorised into the two largest catch-all units of the consensus network, "A" and unassigned ("U"), containing 996 and 1,326 genes respectively.

Further observations include that the male module "C" had two separate consensus modules which it shared highly significant numbers of genes with. It shares 46 with the 46 genes from consensus module "P", and 88 of its 316 genes with the 171 gene consensus module "E". This may be a converse example to male module "Q", whereby a larger module in the male-specific network can be better defined into two

more specific, smaller modules using the consensus expression of 124 as opposed to 66 samples. The concordance of unassigned genes between the two networks is high: 614 of the 691 genes which were unable to be assigned in the male-specific network still had no assignable module in the consensus network, which highlights the replicability between the two networks and that there are certain genes with consistently poor correlations with any others.

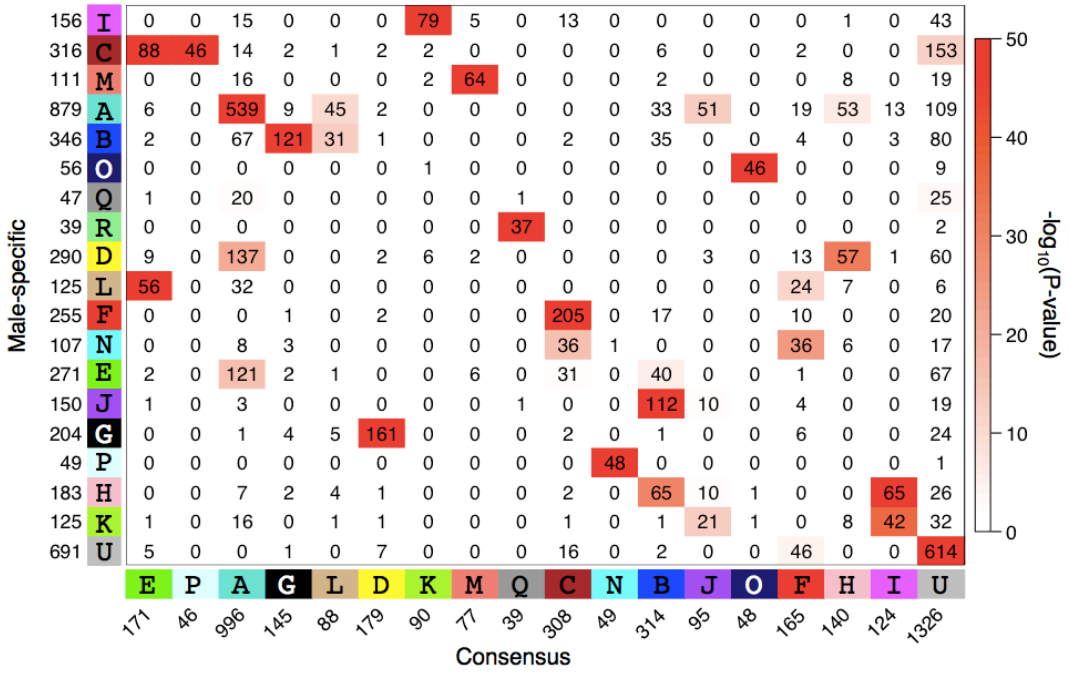


Figure 3.19 Heatmap of gene overlaps between modules from male-specific and male-female-consensus networks.
Numbers in the heatmap correspond to the numbers of shared genes. Cells are coloured by the -$\log_{10}$(p-value) of a Fisher's test for overlap between the corresponding modules.

"Immune system process" was the top hit in GOrilla pathway analysis[355] for the 47 genes from the male-specific module "Q" which had no clearly corresponding module in the consensus network, with a Fisher's enrichment p-value relative to the background expressed genes of 1.00×10^{-7} (1.06×10^{-3} after FDR correction). The 20 genes from module Q which were responsible for producing the enrichment within "Immune system process" are listed in *Table 3.5*. The next 19 most significantly enriched pathways were also all related to immune regulation (*Table 3.6*).

Gene	Description
ITGA4	integrin, alpha 4 (antigen cd49d, alpha 4 subunit of vla 4 - receptor)
IFI16	interferon, gamma - inducible protein 16
CD44	cd44 molecule
FGL2	fibrinogen - like 2
LCP1	lymphocyte cytosolic protein 1 (l-plastin)
DOCK2	dedicator of cytokinesis 2
IFI30	interferon, gamma -inducible protein 30
SAMHD1	sam domain and hd domain 1
NCKAP1L	nck-associated protein 1-like
CD74	cd74 molecule, major histocompatibility complex, class ii invariant chain
IKZF3	ikaros family zinc finger 3 (aiolos)
PTK2B	protein tyrosine kinase 2 beta
CD68	cd68 molecule
PTPRC	protein tyrosine phosphatase, receptor type, c
ETS1	v-ets avian erythroblastosis virus e26 oncogene homolog 1
DOCK10	dedicator of cytokinesis 10
PIK3AP1	phosphoinositide-3-kinase adaptor protein 1
GM2A	gm2 ganglioside activator
TRIM22	tripartite motif containing 22
MAP3K1	mitogen-activated protein kinase kinase kinase 1, e3 ubiquitin protein ligase

Table 3.5 20 genes from male-specific module "Q" which produced enrichment in the GOrilla pathway classification "immune system process"

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0002376	immune system process	1E-7	1.06E-3	3.44 (3844,558,40,20)
GO:0043368	positive T cell selection	1.04E-6	5.52E-3	96.10 (3844,3,40,3)
GO:0045059	positive thymic T cell selection	1.04E-6	3.68E-3	96.10 (3844,3,40,3)
GO:0002252	immune effector process	1.24E-6	3.29E-3	4.34 (3844,310,40,14)
GO:0043383	negative T cell selection	4.15E-6	8.77E-3	72.07 (3844,4,40,3)
GO:0045061	thymic T cell selection	4.15E-6	7.3E-3	72.07 (3844,4,40,3)
GO:0045060	negative thymic T cell selection	4.15E-6	6.26E-3	72.07 (3844,4,40,3)
GO:0046649	lymphocyte activation	5.14E-6	6.79E-3	9.75 (3844,69,40,7)
GO:0042113	B cell activation	6.44E-6	7.56E-3	17.80 (3844,27,40,5)
GO:0030888	regulation of B cell proliferation	6.72E-6	7.1E-3	29.57 (3844,13,40,4)
GO:1903039	positive regulation of leukocyte cell-cell adhesion	6.75E-6	6.49E-3	12.27 (3844,47,40,6)
GO:1903037	regulation of leukocyte cell-cell adhesion	9.03E-6	7.95E-3	8.97 (3844,75,40,7)
GO:0045321	leukocyte activation	1.03E-5	8.34E-3	4.29 (3844,269,40,12)
GO:0022409	positive regulation of cell-cell adhesion	1.38E-5	1.04E-2	10.88 (3844,53,40,6)
GO:0045058	T cell selection	2.04E-5	1.44E-2	48.05 (3844,6,40,3)
GO:0002682	regulation of immune system process	2.67E-5	1.76E-2	3.60 (3844,347,40,13)
GO:0001775	cell activation	3.12E-5	1.94E-2	3.84 (3844,300,40,12)
GO:0030890	positive regulation of B cell proliferation	3.55E-5	2.08E-2	41.19 (3844,7,40,3)
GO:0032944	regulation of mononuclear cell proliferation	4.19E-5	2.33E-2	12.32 (3844,39,40,5)

Table 3.6 GOrilla pathways in which the 47 genes from male-specific module "Q" were significantly enriched

3.4 Discussion

3.4.1 Genome assembly justification

RNA-seq expression data was aligned and quantified against GRCh38 because it represents the newest change in reference sequence since 2009. Approximately 1,000 issues ranging from single base changes to gap-closures have been made since the update from GRCh37, and it was desired to have the most up-to-date possible reference sequence, especially for resolution at the transcript-level and for inclusion of any novel transcripts identified and annotated since the previous release.

3.4.2 Salmon correlation with Cufflinks FPKM

Whilst some skew in *Figure 3.1* may have been attributable to two different quantification units being used, the deviation of the best fit line from the line of y=x demonstrates that there is a trend for certain transcripts to be quantified with a greater expression by Salmon compared to Cufflinks. This trend may be due to the way in which Cufflinks attempts to explain observed expression using the minimum possible number of isoforms required[356]. This parsimonious approach is admirable, though it may sacrifice sequence-specific information contained within reads which can be used by Salmon to increase accuracy of abundance estimates through its bias models[301].

The disagreement in quantifications could also be due to differences in the ways the two algorithms assign effective transcript lengths to reads shorter than the mean read length. In such cases, alignment-dependent algorithms simply use the actual transcript length, whereas alignment-independent tools penalise the likelihood of sequencing fragments shorter than the mean fragment length according to a probability model built from the observed fragment lengths. Additionally, Salmon specifically incorporates bias estimates into the effective transcript lengths as a means of influencing the likelihood of assigning a read to a transcript feature, which could further diverge its approximations from those of Cufflinks. This was demonstrated by Zhang *et al.* who simulated 8 reads from the 100bp long transcript SNGH25-002. Whilst both Salmon and Cufflinks correctly assigned 8 counts to the feature, their estimates of TPM were 20 and 185.6 respectively[356].

Whilst there is no definitive answer for the way effective transcript length should be calculated, the authors then highlight a different example whereby alignment-free

methods clearly do outperform alignment-dependent methods. 154 reads were simulated for the pseudogene RPS28P7-001, with Salmon accurately recapitulating these whilst the alignment-based STAR+Cufflinks approach seriously underestimated its expression and instead assigned almost all reads to the corresponding gene RPS28. The gene RPS28 is alternatively spliced whereas the pseudogene is not. STAR adds bonus alignment scores to spliced reads, precisely in an attempt to penalise spurious alignments to pseudogenes which are assumed to be less highly expressed than their corresponding gene - however in this example and other such cases, alignment-based methods perform poorly compared to alignment-free[356]. This highlights a weakness of alignment-dependent quantification in that there are two separate stages at which technical biases can be introduced - firstly by the idiosyncrasies of the chosen aligner, and then separately by the quantification algorithm - whereas alignment-free algorithms produce results from a single unified workflow.

Salmon quantification was used in this project as opposed to Cufflinks because of its faster speed, lower storage requirements and that incorporation of more bias-correction models. Zhang *et al*.'s same study of quantification methods found Salmon to be more accurate than Cufflinks in a number of metrics when attempting to quantify 50M reads simulated using read distribution statistics from Human Brain Reference RNA sample HBRR-C4 using the RSEM package[357]. Salmon estimated TPMs had a greater Pearson R^2 correlation coefficient than Cufflinks with true simulated values of >0.96 compared to >0.94[356]. Salmon also outperformed Cufflinks when the transcript-content of the gene was more complex. For genes with >15 transcripts, Salmon's R^2 was >0.94 vs >0.92 for Cuflinks. The metric of MARD (Mean Absolute Relative Difference) was also used to quantify differences, calculated as the arithmetic mean of:

$$ARD = \frac{|i - j|}{i + j} \ (for\ i + j \neq 0)$$

where i is the true simulated value and j the value estimated by the quantification algorithms. Again Salmon outperformed Cufflinks by this metric when estimating simulated reads, with a lower overall MARD of 0.170 vs 0.224, and for genes with >15 transcripts of 0.233 vs 0.270[356].

Zhang *et al*. calculated a correlation coefficient of 0.899 between transcript quantification by Cufflinks and Salmon[356]. In addition using Pearson not Spearman correlation, this value is also likely different to the R^2 of 0.678 observed in this study for a number of reasons. Firstly, they performed more stringent filtering before calculating their correlation by removing any transcripts with estimated read counts <5.0. Perhaps more significantly, their correlation was drawn between Cufflinks and Salmon quantifications for a single sample, HBRR-C4, whereas 0.678 was obtained when correlating the median quantification for each transcript across 96 samples by either Cufflinks or Salmon, so there was more potential for variability to be introduced.

3.4.3 Use of Salmon as an alignment-free expression quantification algorithm

Having established that alignment-free quantification consistently outperformed alignment-dependent, Salmon was chosen as the algorithm to use in this thesis.

The similar alignment-free algorithm Kallisto could have been adopted, however there are multiple examples of analyses which have found Salmon to perform better in a range of quantifications based on both simulated and actual RNA-seq data. A 2016 blog by post-doctoral fellow Tom Smith from Oxford University presented one of the first direct comparisons between the recently released Kallisto and Salmon algorithms[358]. Using 100 simulations of random numbers of reads from each of the human protein-coding genes annotated in GRCh38, he demonstrated that Kallisto and Salmon performed very comparably in correlations to ground-truth levels of expression, however Salmon had a higher success rate of assigning zero values to truly absent transcripts than Kallisto across a range of transcripts possessing varying degrees of unique sequence content[358]. This analysis was executed using an earlier version of Salmon, 0.6.0, which did not yet possess the sequence-specific bias model added to version 0.8.0 (the version used in this project), or improvements made to the implementation of the models to reduce risk of over-fitting[301], meaning the advantages of Salmon may have been understated.

The paper introducing Salmon contained comparisons with Kallisto for the quantification of simulations and real RNA-seq samples. When calculating differential expression between 16 RNA-seq libraries simulated to contain realistic GC-bias using the Polyester package[359] and the estimates calculated by the

algorithms, Salmon returned a lower median log fold change of 0.09 compared to 0.14 by Kallisto - when the true change should have been 0.0[301]. Salmon also had a higher sensitivity of 0.409 compared to Kallisto's 0.248 for facilitating downstream detection of truly differentially expressed genes at FDR 0.05 according to simulations[301]. When analysing isoform expression between two sets of the same 15 GEAUVADIS samples sequenced at different centres, Salmon succeeded in ascribing a lower percentage of genes as harbouring switches in the dominantly expressed isoform; only making this mistake for 4.3% of genes compared to 6.8% by Kallisto for genes with at least one transcript expressed at >10 TPM[301].

Zhang *et al*.'s review of RNA-seq quantification also included analyses of Kallisto vs Salmon. Whist Kallisto had leaner memory requirements to quantify 50M 76bp paired-end reads of 3.8G vs 6.6G by Salmon, Salmon was 1 minute quicker taking 6 vs 7 minutes. The Pearson correlation coefficient between their resulting TPMs was high, at 0.966. To probe any specific differences, the authors constructed a test whereby they used the Polyester package[359] to simulate reads from six different transcripts of TP53; the α, β and γ isoforms, and each harbouring the Δ133 variant. The ability to differentiate between such isoforms is of critical biological importance as they can have vastly different effects on tumour progression or suppression, with the full length TP53β isoform inducing apoptosis in cancer cells whilst Δ133β does not[360]. They initially simulated 100 reads for each isoform as a baseline, then to represent different read depths they increased each isoform by 10 and 100-fold. Additionally, to test detection of rare isoforms, they increased the expression of just isoform α by 10 and 100-fold whilst maintaining the expression of the other five at 100 reads. Accuracy was measured by mean MARD scores from 5 replicates of such conditions. One instance where Kallisto did outperform Salmon was in the accurate quantification of the Δ133 isoforms when diluted by isoform α[356]. However, given there were <1 million reads, Salmon could not build bias models during its first streaming phase, and therefore could only estimate biases post-hoc. Despite this limitation, Salmon clearly outperformed Kallisto at quantifying all isoforms in the opposite situation where read depth was simulated to be increased[356]. Given that the libraries analysed in this project were all of very high read depth, Salmon presented the clear choice of an alignment-free quantifier.

3.4.4 Differences between Salmon quantification success rate in primary and cell line samples likely explained by incomplete ribosomal depletion

The success rate for quantification of reads by Salmon was lower than expected for the primary samples from both batches (~50%, *Figure 3.2*), however, the fact that Salmon achieved the expected levels (~80%, *Figure 3.2*) for cell lines indicates that the discrepancy must not be a fault of the algorithm, but of the reads supplied to it generated from the different cDNA libraries

Chromosomes 14, 17 and 21, and the alternative scaffolds of chromosome 21, GL000220.1 and KI270733.1, attracted greater numbers of read alignments by STAR than would have been expected given their lengths (*Figure 3.6*), and the majority of the reads mapping to chromosome 21 and its alternative contigs were those not able to be quantified by Salmon (*Figure 3.6*). The large numbers of reads which were aligned to rRNA regions on these chromosomes (Figure *3.7*) which are not represented by the reference transcriptome likely explains the greater mapping success rate by STAR than Salmon for primary samples (*Figure 3.2*). Accurate quantification of transcript expression is the key requirement for this study though, meaning that Salmon remains the ideal tool from which to identify sQTLs.

STAR mapping of reads to chromosome 21 and alternative scaffolds GL000220.1 and KI270733.1 was less extreme for cell lines than primary samples (*Figure 3.6*). This suggests that excessive ribosomal reads were the cause of poorer than expected Salmon quantification success in the primary samples (*Figure 3.2*), and implies that ribosomal depletion was less successful in primary samples than cell lines. The depletion protocol involves firstly adding DNA oligos with affinity to rRNA sequences followed by incubation with an RNase which digests DNA-RNA hybrids. The incubation step was only carried out for 15 minutes, therefore it is possible that there was incomplete digestion of these sequences. It could also be possible that the digestion was partially successful and created smaller fragments without completely removing them, because the principal reason for STAR mapping failures was reads being too short (*Figure 3.3* d). Fragments would have been size-selected prior to sequencing, but given the probabilistic nature of the selection process, some smaller fragments may have been included.

The manufacturers of the ribosomal depletion kit, New England BioLabs, were contacted about the apparent failure of the ribosomal depletion. They suggested a number of steps at which the depletion could have been sub-optimal:

- If the RNase did not fully digest the hybridized rRNA then it would have persisted through to sequencing.
- The kit contains a DNase to digest any unused ssDNA probes which hadn't bound to rRNA, however if this enzyme failed then the probes would be sequenced and could appear as rRNA reads.
- It is also possible for the kits to be overloaded, and if the primary samples contained an excess of rRNA then it could have exhausted the reagents and persisted through to sequencing.

It could be that the difference in performance was because the protocol and kit for ribosomal depletion was developed using cell lines, not primary tissue samples, and therefore is in some way optimised to their specific transcriptomic profile - because PCA clearly demonstrated a difference between tissue samples and cell lines (*Figure 3.10*). Or there could be an intrinsic quality of the mucosa-derived primary samples which makes rRNA depletion more challenging.

Zhao *et al.* compared expression profiles obtained from RNA-seq using either poly-A enrichment or ribosomal depletion from 11 fresh frozen primary breast tumour samples. Poly-A enrichment produced 62.3% of reads mappable to the transcriptome by the genomic aligner MapSplice[361], whereas the two different ribosomal depletion protocols they used (either Ribo-Zero-Seq or DSN-Seq) only produced 31.5% and 22.7% of reads which could be mapped to the transcriptome[316]. This demonstrates that other groups have also observed low proportions of transcriptomic reads after ribosomal depletion.

Lahens *et al.* used *in vitro* transcription of >1,000 cDNAs with known abundance and sequence content to investigate the technical sources of noise and bias in RNA-seq protocols. They found that rRNA depletion accounted for the most significant variability in sequence coverage across all transcripts[362], further highlighting the potential of this stage of the RNA-seq protocol to influence the outcomes of sequencing.

In cell lines Salmon had a higher quantification success rate than STAR (*Figure 3.2*). Perhaps this is because for a library where rRNA depletion has been more successful, the multiple bias-detection models constructed by Salmon allow it to outperform traditional alignment algorithms. A 2-pass STAR alignment may have produced even higher mapping successes, however previous in-house analyses showed the additional benefit over 1-pass to be minimal. The absence of large amounts of alignments to repetitive rRNA sequences at chromosome 21 or its associated contigs in cell lines (Figure *3.7*) likely contributed to Salmon achieving a greater percentage quantification success rate in these samples.

The rRNA sequences were not explicitly removed prior to running sQTL detection algorithms, however the packages will have automatically disregarded such sequences because each rRNA gene only possesses a single annotated transcript, therefore an sQTL cannot be called for them.

3.4.5 Differences in mapping success rate between batches likely explained by total read depth

There was a greater depth of sequencing for samples from batch 10525 (with which the cell lines were also sequenced) than batch 2013152, with means of ~155M and ~130M reads respectively (*Figure 3.2*). A greater percentage of the reads not quantifiable by Salmon were able to be successfully mapped by STAR for batch 2013152 (~90%) than from 10525 (~65%, *Figure 3.3* a). The percentage of reads which failed STAR mapping for being too short was lower in batch 2013152 (~10%) than 10525 (~30%, *Figure 3.3*, b). This implies that the extra sequencing depth in batch 10525 resulted in the sequencing of more poorly mapping sequences than in batch 2013152, as the likelihood of rare fragments being sequenced does not scale exactly linearly with sequencing depth[363]. Another consequence of using ribosomal depletion as opposed to poly-A-enrichment is the potential presence of immature reads from intergenic regions. With greater sequencing depth of an rRNA depleted library, such noisy, low-frequency transcription could be the cause of the poorer percentage mapping success by STAR in batch 10525.

3.4.6 PCA and batch effects

There was clear separation of primary and cell line samples by PCA (*Figure 3.10*), which is to be expected given that the primary samples were normal mucosa and

the cell lines were immortalised CRC lines (HCT116, SW480 and LS174T). It was initially surprising that there was no gender separation by PCA of the 221 patient samples from both batches (*Figure 3.11*). However, it became clear that the differences were being masked by batch effects and the contribution of low variance features, because the expected separation became apparent when PCA was performed on just the highest variance genes from a single batch (*Figure 3.13*).

Correction with ComBat did coalesce the two batches together to some extent (*Figure 3.14*), however it introduced too many negative numbers in to the data for it to have been compatible with the sQTL detection algorithms used in this project (*Figure 3.15*). The hidden factors identified by PEER analysis were unable to be used with sQTLseekeR due to its inability to accept covariates, and the residuals also contained too many negative values (*Figure 3.16*).

These procedures were followed in an attempt to remove batch effects and allow the combining of the two batches for subsequent sQTL analysis. However, the authors of the sQTLseekeR package make the point that because it uses relative transcript ratios to call sQTLs, common batch effect confounders such as total library size and library content should be considerably ameliorated[364]. Taking ratios of transcript expression will internally normalise each gene's transcript content within each sample, meaning that changes in relative transcript ratios between samples of different genotypes will still be detectable regardless of batch differences such as library size, as long as there are sufficient read counts of each transcript involved in the sQTL for it to be considered statistically significant[364]. A similar appraisal could be made of the relative changes in PSI used by the Leafcutter algorithm[261], however the FastQTL association tool used in concert with it is able to accept covariates[259], so both approaches are covered in this thesis.

3.4.7 WGCNA differential network expression analysis

The Fisher's exact test p-values for comparing gene membership between modules were not multiple testing corrected by the WGCNA software before plotting the heatmap in *Figure 3.19*. A stringent Bonferroni correction would have required their values to be multiplied by the total number of comparisons made between modules (342 separate pairwise comparisons for an 18 by 19 matrix), and so any p-values < 1.46E-4 would have achieved multiple-testing corrected significance. Given that the -log10(p-value) scale for *Figure 3.19* ranged from 0 to 50, the majority of

comparisons with any visible colouration would have remained statistically significant after such a correction.

The majority of gene co-expression modules identified by WGCNA remained stable between male-specific and consensus gender networks (*Figure 3.19*). One male-specific module, the genes of which were enriched in immune-related functions, was not re-created in the consensus network (*Table 3.5*). A similar level of agreement between gender expression networks was observed by Fatima *et al.* when performing WGCNA on 44 male and 42 female peripheral blood mononuclear cell (PBMC) samples. They constructed differential networks for each gender, comparing RNA-seq from 0 or 240 minute timepoints following oral lipid treatment in an attempt to find biomarkers for early-stage diet-related diseases. They found clear similarities between the response of the two genders, with three of five genes differentially expressed in females being the only three genes differentially expressed in males, and the gene module with the most significantly different regulation between the 0 and 240 minute networks relating to G-protein coupled receptor activity in both genders[365]. There were only two modules which were specific to females, which related to energy metabolism and the innate immune inflammatory response[365].

Given the minimal differences observed between gene expression networks, it was considered justified to include both male and female samples together in the same sQTL analyses in this project.

Chapter 4 Generation of sQTLs

4.1 Introduction

This chapter focuses on the identification of sQTLs via two separate software packages. The significances, effect sizes and classes of sQTL events are analysed, in addition to their local distributions relative to the features they correspond to, and their genome-wide distributions. The correspondence between the two algorithms is explored, and thresholds are set to filter out low effect size events prior to functional characterisation in the following results chapter.

4.1.1 Choice of sQTL detection algorithms

Using short read alignment to quantify changes in transcript expression is challenging due to the high sequence similarity that many transcripts share with each other. Therefore to detect as comprehensive a list of sQTLs as possible, two complementary algorithms were used: one which works at the transcript-level and one which uses PSI. A similar approach was taken by the GTEx Consortium, which used sQTLseekeR and Altrans for their 2015 analysis of 9 tissues[200].

A more recently published software package, Leafcutter, outperforms Altrans in sQTL identification. When analysing the same population of 372 GEAUVADIS LCLs, Leafcutter was able to identify 1,294 and 1,982 sQTLs at 1% and 5% FDR respectively, in comparison to 624 and 1,083 by Altrans[254,260]. Leafcutter is based on the concept of quantifying differential intron usage, rather than exon linkages as in Altrans, and has greater potential for identifying novel events than Altrans as a result of its ability to infer the presence of introns *de novo* as opposed to requiring a reference genome to define exon boundaries.

sQTLseekeR is more nuanced than simpler measures such as transcript ratio QTLs (trQTLs), because it requires a reciprocal change between two transcripts of the same gene and takes into account the interdependence of all transcripts within a gene. DRIMSeq is another algorithm which takes a multivariate approach to sQTL identification; however sQTLseekeR has been demonstrated to identify more sQTLs than DRIMSeq from the same dataset. When analysing GEAUVADIS CEU LCLs, sQTLseekeR identified 3,699 significant SNP-to-gene associations for 191 genes at FDR 0.05, compared to 3,036 in 97 genes by DRIMSeq[246]. From YRI LCLs sQTLseekeR identified 3,449 significant associations in 258 genes compared to

1,867 in 63 by DRIMSeq[246]. sQTLseekeR proved more adept at identifying sQTLs in genes with lower expression and fewer constituent transcripts expressed[246]. The improvement in performance should not be due to false positives from high transcript ratio variances because sQTLseekeR has the ability to discount candidate sQTLs where relative transcript expression variance is different between genotypes (svQTLs)[247].

sQTLseeker and Leafcutter have unique advantages and limitations which make them theoretically more adept at identifying different classes of splicing event. Whilst sQTLseekeR requires a known reference transcriptome, it is able to identify complex alternative splicing events capturing changes in expression of multiple exons and UTR regions. Leafcutter can identify sQTLs involving the excision of novel introns, but is less adept at identifying complex events involving multiple features, and cannot identify changes in 5' or 3' UTRs which do not involve a change in intron structure.

4.2 Methods

4.2.1 Genotyping

Individuals of the Scottish Colorectal Cancer Susceptibility and the Colorectal Cancer Genetic Susceptibility (SOCCS or COGS) cohorts (batch 2013152) were genotyped on the Illumina HumanOmni5M-4v1_B SNP-Chip, which detects 4,327,109 variants. Individuals from the Scottish Vitamin D study (SCOVIDS) cohort (batch 10525) were genotyped on the Illumina OmniExpressExome BeadChip 8v1.250 (OEE3), which detects over 270,000 exonic SNPs. Images were obtained using the HiScan H166 scanner and .cel files were quantified using GenomeStudio v2011.1. Both arrays were imputed to the 1000 Genomes Phase 3 release[366] based on GRCh37.p13, which generated a total of 47,246,411 SNPs. SNPs common to the two datasets were subjected to QC thresholds of 5% MAF and 0.8 imputation quality score, which left 5,917,734[4]. The genotype coordinates were lifted-over to GRCh38.p10 using the UCSC Liftover Tool[367] with chain file "hg19ToHg38.over.chain.gz". Five individuals were discounted from the study for presenting as clear outliers in a PCA of genotypes obtained from the OEE3 SNP array (*Figure 4.1*)[i]. It should be noted that the apparent separation between two clusters in the bottom right quadrant of the plot was not attributable to either gender

[4] Genotyping and imputation analysis performed by Dr Maria Timofeeva

or batch, and was considered minor enough to submit all such individuals to the same analysis so as to increase the power to identify sQTLs by including as many samples of expression data as possible.

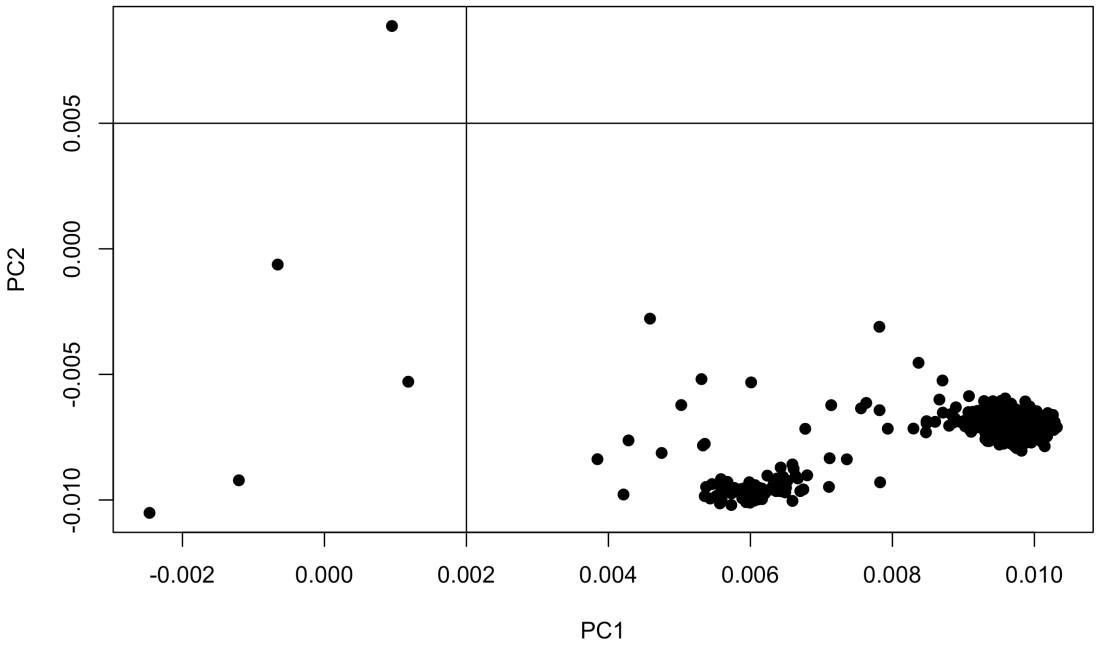


Figure 4.1 Principal Components Analysis of samples genotyped on OEE3 SNP array showing 5 individuals which were excluded from future analysis[5].

Where genotypes were in decimal dosage form, they were rounded to the nearest integer; either 0, 1 or 2. 3,309 variants with dosages of exactly 0.5 or 1.5 were removed as they could not be definitively assigned to either genotype. Variants with <5 individuals in each of the three genotype groups (0, 1 or 2) across the cohort of 221 samples were removed, reducing the number of SNPs to 3,978,682. The approach recommended by the sQTLseekeR authors, and adopted by the GTEx consortium, was to only require ≥2 of the 3 genotype groups to contain ≥5 individuals[246,247]. With a desire to be more stringent in this study, all 3 genotype groups for a given variant were required to contain ≥5 individuals in order to be included, because preliminary analyses with more lenient thresholds suggested that <5 individuals did not constitute sufficient data to generate reliable mean transcript expression ratios.

INDEL status of the CRC risk locus 30kbp upstream of SHROOM2 previously identified by the CCG Group[98] was assessed by either WGS or amplicon

[5] This principal components analysis was performed by Maria Timofeeva.

sequencing. Primers were developed which produced a 184bp amplicon if the 24bp sequence was present and 160bp if it was missing. The reverse primer contained a fluorescent 6Fam tag on the 5' end allowing PCR products to be detected on an ABI microfluidics machine and analysed via GeneScan software to produce accurate sizing[6]. For the 152 samples for which there were both WGS and amplicon performed, 149 agreed and 3 disagreed, with the variant not analysed in any samples for which there was a conflict.

Forward Primer: CACCCACATCCCGCTGATTG

Reverse Primer: CCTTACCAAGAGGCGAA

4.2.2 sQTLseekeR

There were clear batch effects in the RNA-seq data, as detailed in the previous chapter. Batch correction was attempted using two separate methods, though each produced negative values incompatible with the sQTLseekeR algorithm. The sQTLseekeR authors posit that taking transcript expression as a ratio of the total expression of the gene from which the transcript originates should internally control transcript expression for each individual. Switches in relative transcript expression between samples of different genotypes should still be detectable, regardless of differences between batches in e.g. library size, read length or insert size distribution. They suggest that the most unbiased approach is to simply provide raw transcript counts to the algorithm, rather than RPKMs or TPMs, which can be influenced by factors such as library size.

Following the approach taken by the GEAUVADIS Consortium[254] and the GTEx Consortium[200], only protein-coding and lncRNA genes were used for sQTL detection in order to reduce the multiple testing burden that would be incurred if other gene classes of less interest to this study were included, such as pseudogenes or various classes of micro RNAs.

sQTLseekeR filters expression data prior to calculating transcript expression ratios, removing transcripts with low expression by setting a default threshold of requiring a transcript to have read counts of above 0.01 across all samples. Decimal counts are applicable because Salmon is able to assign probabilistic expression values to

[6] Genotyping lab work carried out by Stuart Reid and Marion Walker, WGS by Victoria Svinti.

features which need not be integers. Any genes with only one remaining transcript passing expression thresholds were excluded.

Transcript expression ratio is calculated by dividing each transcript's expression by the sum of all transcripts for the gene. Genes with dispersion for transcript expression ratio of less than 0.1 were removed because low variability in expression between samples would make a gene unlikely to yield a significant reciprocal change in transcript expression between genotype groups. The default settings for expression filtering were used to ensure comparability with other studies employing sQTLseekeR[200,247].

When searching for SNPs to associate with transcriptional changes, sQTLseekeR sets a window consisting of the entirety of the gene body +/- a custom search window up and downstream of the gene. The authors recommend a default window of 5kbp as they make the assertion that SNPs most likely to be causal of transcriptional changes will be within or in close proximity to the transcript sequence in question. The same window was adopted in this study for the purposes of consistency and comparability[200,247]. FDR correction was run via a built-in function from the sQTLseekeR package applying a Storey q-value[368] threshold of 0.05. A more stringent threshold of 0.01 was applied for the detection of svQTL events, and any candidate svQTLs were removed from the resultant pool of sQTL events.

sQTLseekeR defines the effect size of an sQTL event as the maximum difference (MD) in relative transcript expression observed between two genotype groups. For the purposes of certain analyses in this chapter, it was desirable to identify the "lead" sQTL SNP for a given event, e.g. for a particular pair of reciprocally changing transcripts. In such an instance, the most significantly associated SNP per transcript-pair was chosen based on the lowest Storey q-value, then if any were tied by significance, the SNPs corresponding to the largest MD changes were retained. Any further ties were broken arbitrarily by choosing the first row of the remaining data. Approximately 25% of sQTLs had >1 SNPs which were equally tied for significance and effect size in this way, implying they were in high LD with each other.

In addition to the 3.9M SNPs which passed filtering, sQTLseekeR was also run using the INDEL status of SHROOM2 encoded in the form 0, 1 or 2, with 0 denoting no copies of the 24bp.

4.2.3 Leafcutter data preparation

Prior to input to Leafcutter, the RNA-seq fastq files were aligned to the GRCh38.p10 reference genome with the Ensembl v88 gene build using STAR[302] 1-pass alignment with default parameters[8]. A threshold was applied that intron clusters must be supported by at least 50 reads across the cohort of 221 samples, and a given intron cluster must have reads supporting it in at least 40% of the samples. Introns of length up to 500kbp can be inferred by Leafcutter.

Intron excision ratios are calculated by dividing the number of reads supporting each intron by the sum of all reads supporting all introns in the corresponding intron cluster. Intron excision ratios were zero-centred and quantile normalised (using the Python *scipy.stats* package[369]) to make them compatible with the model implemented in the FastQTL package[259], which is used to calculate associations between intron excision ratios and variants. Peaks in the distributions of raw intron excision ratios towards 0.0 and 1.0 indicate that the majority of introns are either rarely excised or are commonly excised (*Figure 4.2* a). There is a shallow peak around 0.5 indicating introns which are included approximately 50% of the time. The process of normalisation then results in a clear deviation from the raw distribution of intron excision ratios (*Figure 4.2* b). However, the authors justify the process through citation of a previous study of DNase QTLs, which demonstrated that normalisation of feature data increased the number of identifiable QTL associations[370]. A filter removing rows in the bottom 2% of variance was applied because these introns would be unlikely to yield alternative splicing QTL events.

[8] STAR alignment performed by Dr Alison Meynert

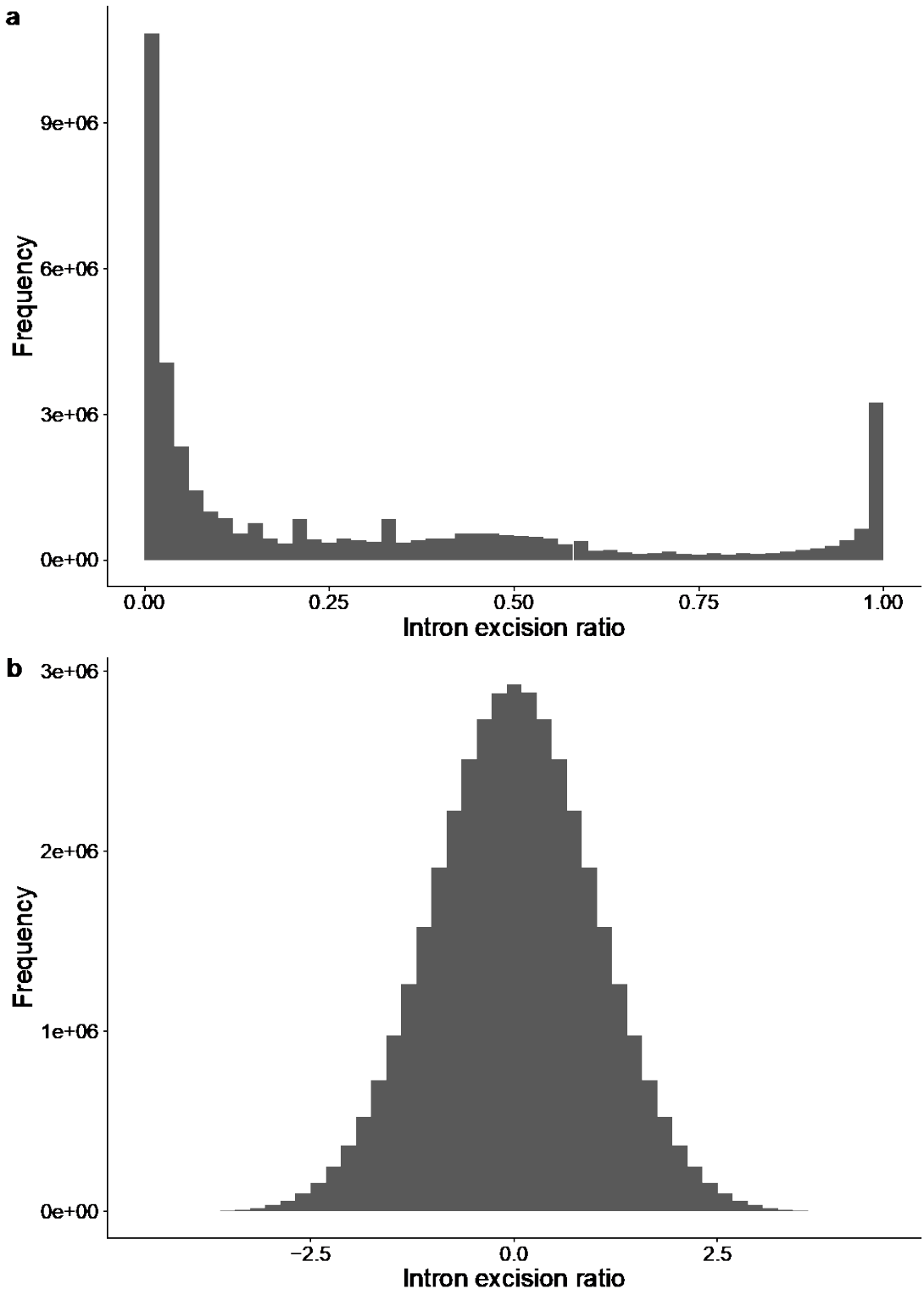


Figure 4.2 a) Raw intron excision ratios from all chromosomes of 221 patients as quantified by Leafcutter.
b) Intron excision ratios after zero-centring and quantile normalisation.

4.2.4 sQTL associations with FastQTL

Leafcutter is able to prepare matrices of intron excision ratios, but it requires a separate program to calculate correlations between these phenotypes and genetic variants. The Leafcutter authors recommend FastQTL[259]. For each feature, after calculating nominal p-values for associations between all SNPs within the search window of the feature, permutations of phenotypes between individuals are used to model the tail of a null distribution positing no association between the feature and a given variant. The tail of the distribution is what determines the corrected significance ascribed to an association, and it can be modelled with relatively few permutations accurately enough via a beta distribution to calculate adjusted P-values efficiently for analyses where many millions of associations need to be tested. FastQTL was run with a minimum of 1,000 permutations and a maximum of 10,000 to construct the null distribution per association (the process can be halted early by the algorithm to save computational resource if a plateau is reached whereby no change in predicted parameters occurs with increasing permutations). *Figure 4.3* demonstrates strong concordance between the p-values estimated via the beta-distribution and empirical p-values, indicating the estimation process worked as expected.

The beta-estimated p-values were corrected for the number of SNP associations tested in *cis* with a given intron, with only the most significant SNP associated with each feature returned by FastQTL. The p-values were then Bonferroni corrected[371] for the number of introns tested per intron cluster, because the usage of a given intron is not independent of other introns which share the same boundaries. Then, to reach full genome-wide significance, all resulting p-values were subjected to Benjamini-Hockberg[372] correction with a threshold of 0.05, following the methods implemented by the Leafcutter authors in their 2018 paper[260].

The first 9 principal components of the intron excision ratios as calculated by Leafcutter were supplied as covariates to FastQTL (*Figure 4.4*).

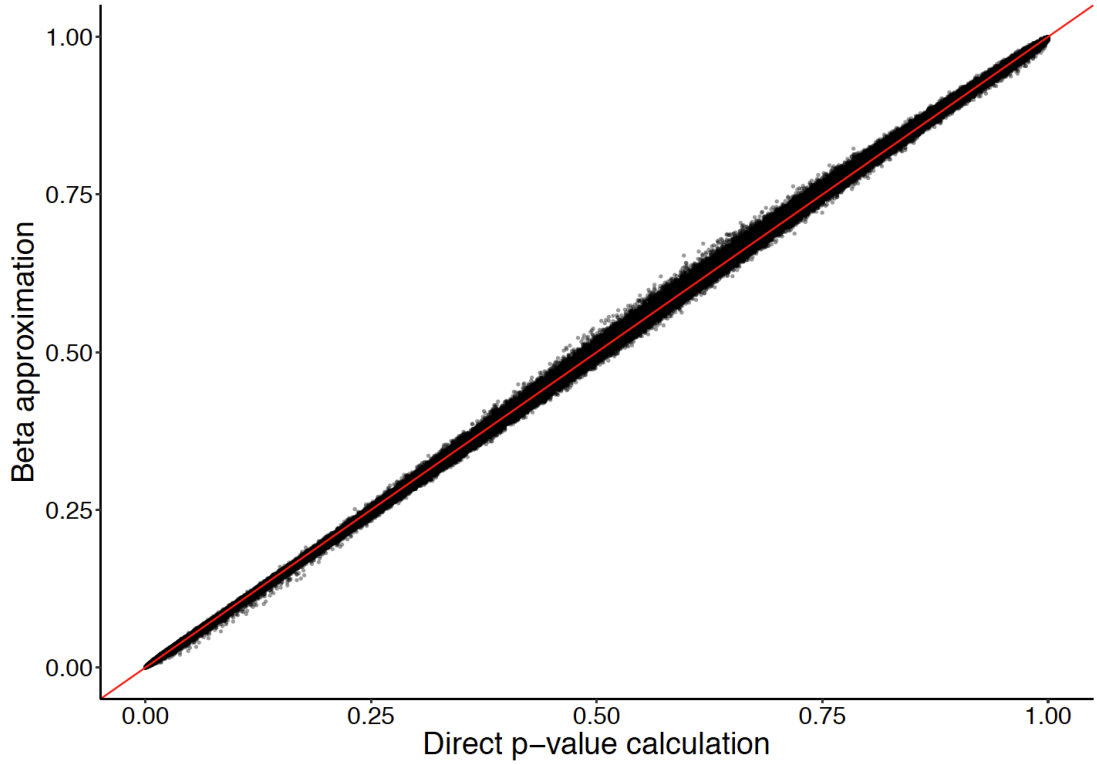


Figure 4.3 Correlation between beta-approximated and empirical p-values generated by FastQTL. Red line indicates a vector of y=x.

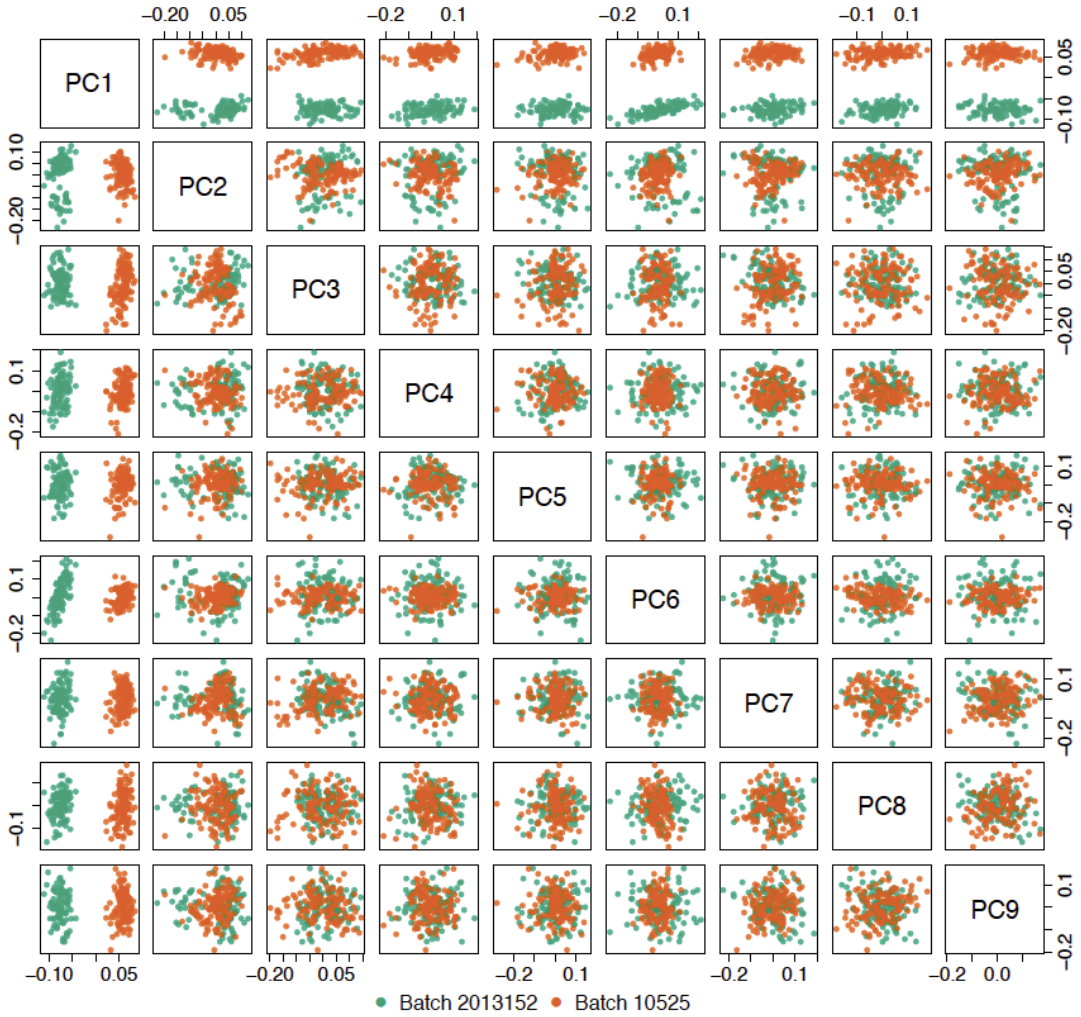


Figure 4.4 Principal components of Leafcutter intron excision ratios
Samples coloured by sequencing batch, which shows separation in the first principal component.

The effect size of sQTL events for Leafcutter are calculated as twice the absolute value of the slope of the FastQTL linear association between intron excision ratio and genotype.

One limitation of FastQTL is that when associating variants in *cis*, it only sets a search window around the single coordinate indicating the start of the feature being investigated: in this case the start coordinate of introns inferred by Leafcutter. The search window was set to 100kbp in line with that used by the authors' 2018 methods paper[260]. A larger window than for sQTLseekeR is required in order to increase the likelihood of capturing the entirety of the gene body - which sQTLseekeR does automatically but Leafcutter does not. 93.3% of protein-coding and lncRNA genes have a total length of ≤190kbp, as annotated in the Ensembl

gene build v88[328], which means that a search window of ±100kbp would be capable of surveying their entire length plus an additional 5kbp at either end, as sQTLseekeR adds. Leafcutter can infer introns up to 500kbp long, which doesn't necessarily tally with a search window of only 100kbp; however features this large were extremely rare, only 1,908 introns inferred by Leafcutter (1.104%) had a length ≥100kbp, and only 30 introns (0.017%) had a length ≥400kbp. It would have been overly stringent to penalise the inclusion of introns of this length when relevant sQTL SNPs could still have been discovered within ±100kbp of their start coordinates.

A further caveat is that the intron boundary coordinates inferred by Leafcutter are unstranded, because they are agnostic of gene build. The smallest genomic coordinate is always annotated as the intron "start", around which the search window is extended. However, in reality, 50% of the time the search window will be extended from the end coordinate of an intron relative to the direction its gene of origin is oriented genomically, given that there is an approximately equal division of genes residing on each strand.

4.2.5 Filtering of sQTL events

It is possible that despite FDR correction some of the identified sQTLs may represent false positive changes in transcript expression or intron excision due to low effect sizes or low total expression. Therefore post-hoc thresholds were applied to these parameters to refine the sQTL list to the most likely functionally relevant candidates. Where necessary, the lists of filtered sQTLs were compared to all significant sQTLs to assess whether they contained larger proportions of putative functionally relevant variants.

Different strategies were attempted to identify suitable effect size thresholds from the data in an unsupervised way. These included: attempting to identify the effect size threshold which produced the highest agreement of genes containing sQTL events between the two packages; choosing an effect size threshold which maximised the number of CRC-relevant GWAS gene sQTLs; or increasing the threshold until significant and non-significant events formed separate distributions in the effect-size dimension. However, none of those approaches yielded a consistent or clear threshold.

The authors of the sQTLseekeR package arbitrarily designate an MD value of ≥0.2 as a threshold above which changes in transcript level can be considered

biologically relevant. Adopting this threshold retained the top 9.29% of sQTLseekeR events (*Table 4.1*). In order to apply a comparable threshold to Leafcutter events, the same percentage of events with the largest absolute effect sizes were retained. The corresponding 2*absolute slope value which imposed the same percentage was 2.25284 (*Table 4.1*).

A post-hoc threshold was also set for the gene expression or intron-usage counts from which sQTLs could be drawn, so as to increase the confidence in any sQTLs identified. *Figure 4.18* demonstrates that sQTLs associated with lowly expressed genes are outliers in the effect size distribution, which contributes to a negative correlation between expression and effect size of sQTLs. Therefore an expression threshold was set requiring sQTLseekeR-derived events to have emanated from genes with a mean log10(count) ≥ 1.0 across all 221 samples. This retained the top 7.79% of events (*Table 4.1*).

A post-hoc filter was also applied to Leafcutter sQTLs, given a similar bias from low expression events - though which caused a correlation in the opposite direction (*Figure 4.19*). This filter was set at a less stringent level of $\log_{10}(\text{count}) \geq 0.5$ because Leafcutter identifies events at the level of individual introns rather than whole transcripts, meaning that fewer counts are typically available to support each intron than each transcript. This retained the top 8.37% of Leafcutter events (*Table 4.1*).

seekeR MD threshold	Leafcutter 2*abs(slope) threshold	seekeR log10(Gene counts)	Leafcutter log10(intron counts)	seekeR % events retained	Leafcutter % events retained
0.20	2.25284	0.0	0.0	9.29	9.28
		0.5	0.5	8.87	**8.37**
		1.0	1.0	**7.79**	5.99
		1.5	1.5	6.54	3.52
0.15	1.98081	0.0	0.0	14.8	14.8
		0.5	0.5	14.4	12.6
		1.0	1.0	13.0	8.78
		1.5	1.5	11.1	5.24
0.10	1.626546	0.0	0.0	26.2	26.2
		0.5	0.5	25.8	20.8
		1.0	1.0	24.3	13.9
		1.5	1.5	21.5	8.29

Table 4.1 Thresholds applied to sQTLseekeR and Leafcutter events and the percentage of events retained.
The combinations of thresholds chosen to be applied to each package are underlined. sQTLseekeR shortened to "seekeR" in column headers.

4.3 Results

4.3.1 Distribution of sQTLseekeR events

sQTLseekeR identified 97,021 significant sQTL SNPs associated with transcript switching events in 3,420 different genes at genome wide FDR 0.05. The median number of sQTLs identified per gene was 10 and the mean 28. The distribution is skewed by 55 genes for which there were >200 significantly associated SNPs (*Figure 4.5*a). These 55 highly associated genes spanned 17 different chromosomes and were not solely dominated by genes in the MHC region of chromosome 6, as might have been expected due to its dense polymorphisms[373,374]. The highest associated HLA genes from the MHC region were HLA-B with 628 separate significant sQTLs, HLA-DPA1 with 204 and HLA-DQB1 with 109.

Within the 3,420 different genes with sQTL events, there were 5,492 different pairs of transcripts which underwent reciprocal changes in expression ratios (*Figure 4.5*b). The distribution of the number of SNPs associated with each transcript-pair was similar to that of the SNPs per gene (*Figure 4.5*c).

There was a predominance towards lower effect-size (MD) events identified by sQTLseekeR, and no clear correlation between MD and significance (*Figure 4.6*).

The “banding” of p-values is caused by limitations of the non-parametric MANOVA-based test that underlies sQTLseekeR, which reaches a maximum significance it can ascribe to an event given the number of samples tested in this analysis. There was also no apparent difference between the relationship of these values for protein-coding transcript-switches which occurred between transcripts with the same or different biotypes, or for transcript-switches that occurred between protein-coding or lncRNA genes. The summary of all biotype changes is presented in *Table 4.2*.

Transcript Biotype Change	Number	Percentage
protein-coding to protein-coding	2776	50.546
protein-coding to retained_intron	839	15.277
protein-coding to processed_transcript	639	11.635
protein-coding to nonsense_mediated_decay	529	9.632
protein-coding to non_stop_decay	2	0.036
protein-coding to other	396	7.211
lncRNA to lncRNA	310	5.645
lncRNA to retained_intron	1	0.018

Table 4.2 sQTLseekeR transcript biotype changes

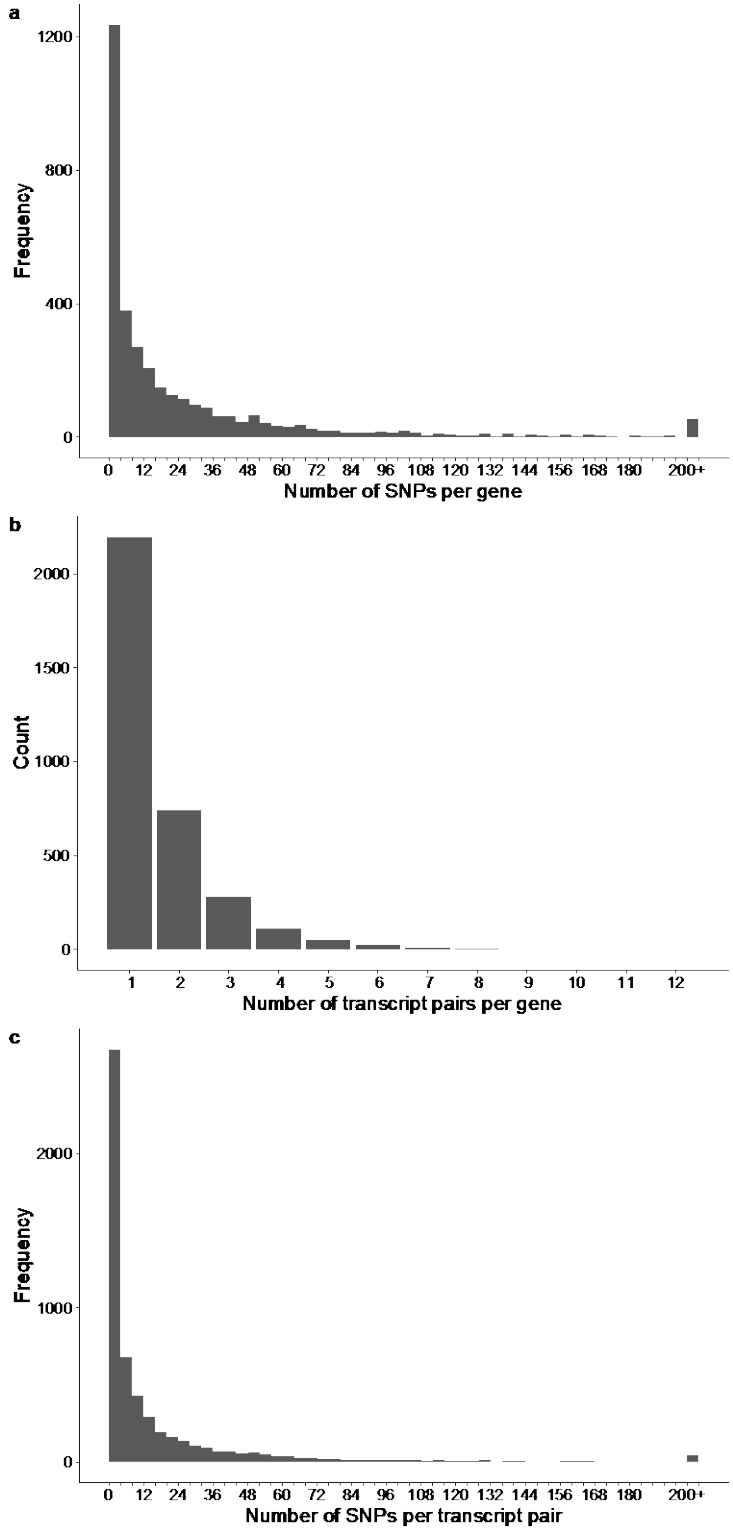


Figure 4.5 a) Number of sQTL SNPs significantly associated with each gene b) Number of significant transcript-pair switches per gene. The modal number of significant transcript-pair switches per gene at FDR 0.05 was 1, the mean was 1.61 **c) Number of sQTL SNPs significantly associated with each transcript pair**. The median number was 5.00 and the mean 17.67. Any genes or transcript-pairs with ≥200 associated SNPs are binned for clarity

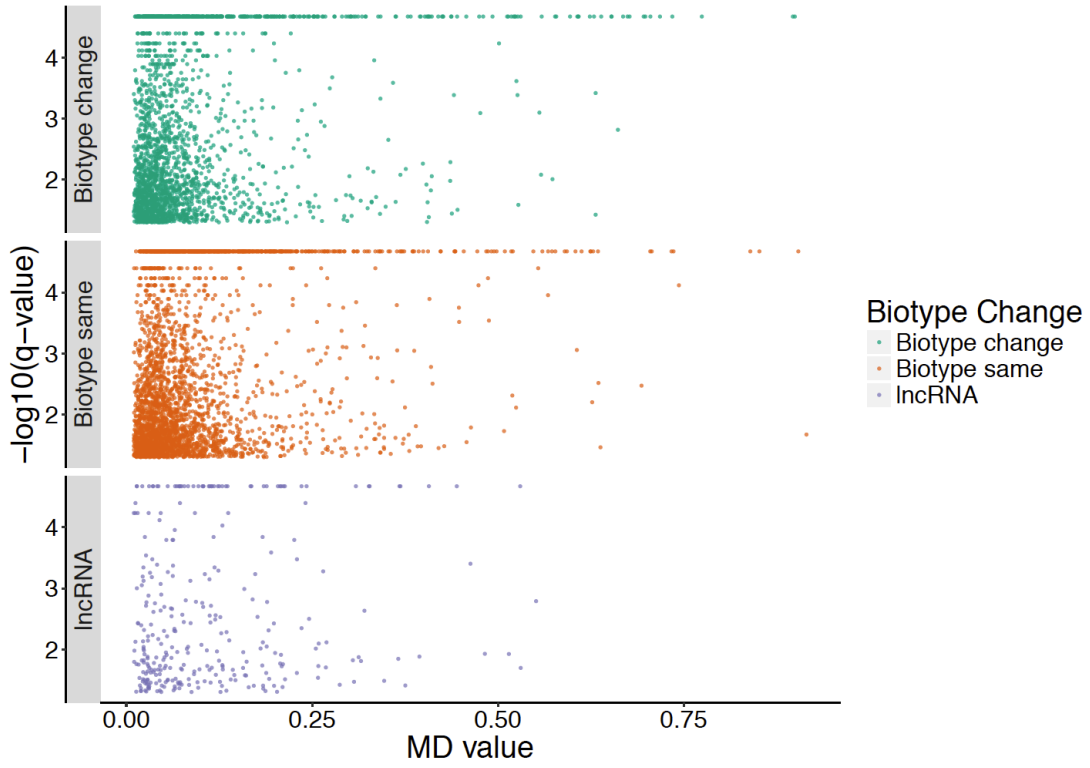


Figure 4.6 Distribution of significance of sQTL events against MD value
For all 5,492 FDR 0.05 significant transcript-pair switches, the significance of the event expressed as -$\log_{10}$(Storey q-value) is plotted against the MD effect size of the event. Transcript-pairs are faceted by whether they were from protein-coding or lncRNA genes, and the protein-coding events are further separated by whether there was a change in biotype between the two transcripts involved in the event or not. Protein-coding with biotype change (n = 2205), Protein-coding with no biotype change (n = 2976), lncRNA (n = 311).

4.3.2 Classification of sQTLseekeR Splicing Events

Sammeth *et al.* developed an ontology for describing alternative splicing events[375] which formed the basis of the AStalavista package[376], and more recently has been incorporated into the sQTLseekeR package. The most common class of sQTLs identified by sQTLseekeR are complex events (*Figure 4.7*). Most types of event are equally represented whether analysing the 3,420 most significant sQTLs per gene, or whether analysing the 5,492 most significant sQTLs per pair of transcripts. This implies that where multiple transcript-pair events are found per gene, they do not constitute systematically different classes of splicing events to the single per-gene events.

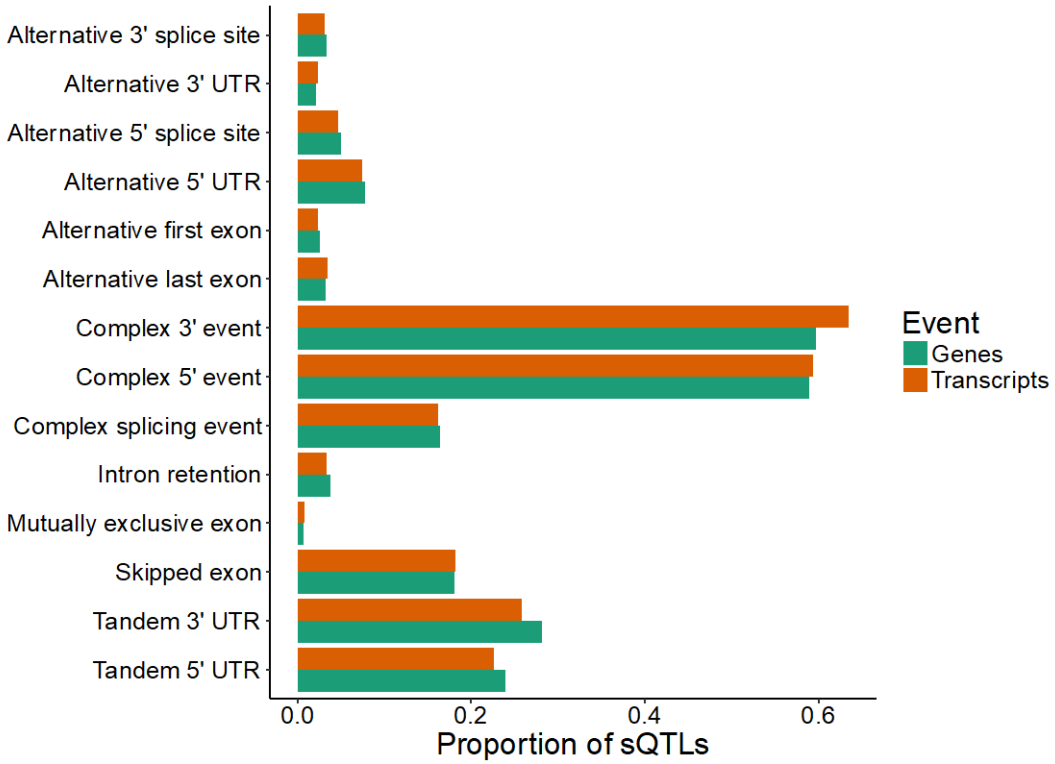


Figure 4.7 Classification of sQTL splicing events identified by sQTLseekeR
The proportion of sQTLs which are able to be classified as each type of splicing event sum to greater than 1.0 because not all classes are mutually exclusive and some events can satisfy the criteria for more than one class.

4.3.3 Distribution of Leafcutter Events

Leafcutter inferred the presence of 175,792 different introns belonging to 47,977 separate “intron clusters”. After association with variants via FastQTL, there were 12,830 significant intron-level sQTL events from 6,153 intron clusters at FDR 0.05 (*Figure 4.8*). Plotting effect sizes against significance for Leafcutter sQTL events more closely resembles the expected pattern of a volcano plot than was observed for the non-parametric sQTLseekeR, with larger effect size events trending towards greater significance (*Figure 4.9*). Because FastQTL calculates the significance of associations using a parametric linear model, there is no p-value banding as seen with sQTLseekeR.

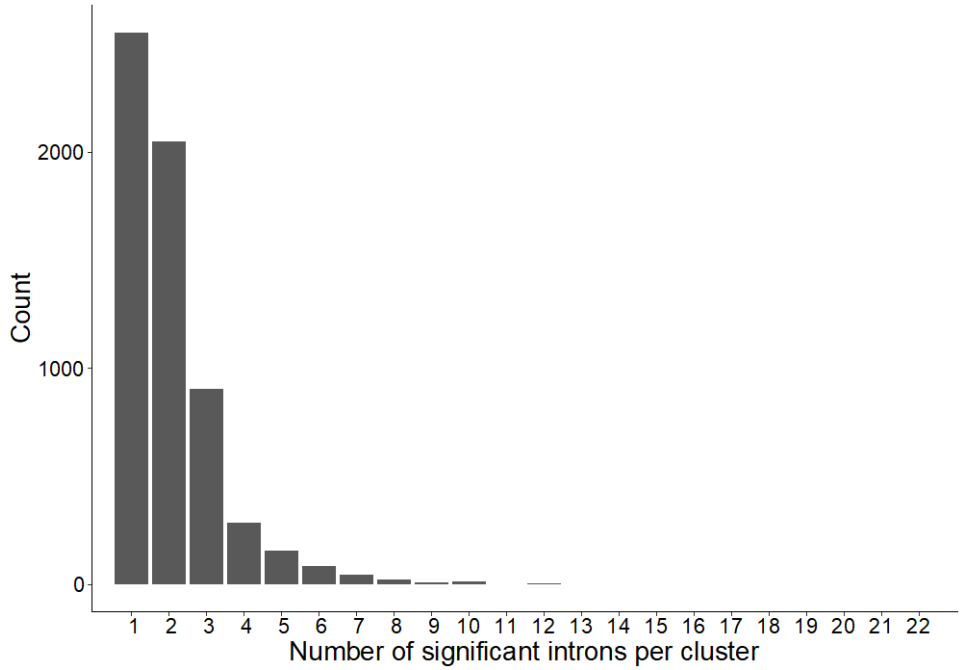


Figure 4.8 Significant introns per intron cluster.
The modal number of significant introns per cluster was 1.00, the median 2.00 and mean 2.09. One cluster with 46 significant introns has been excluded for clarity.

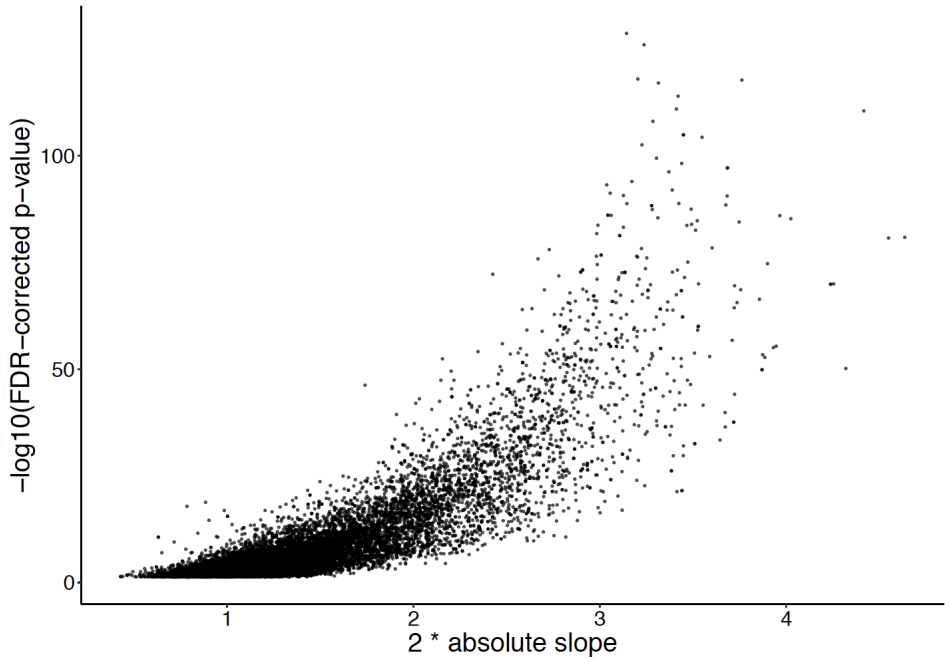


Figure 4.9 Leafcutter significance against effect size

4.3.4 Assignment of Leafcutter events to genes and transcripts

There is no direct way to link an intron-level sQTL event identified by Leafcutter to an Ensembl-annotated gene or transcript, because of the intrinsic ability of Leafcutter to identify introns agnostically of a reference transcriptome.

In order to identify the features which the Leafcutter-inferred introns most likely correspond to, introns were firstly assigned to any genes annotated in Ensembl gene build v88 for which their coordinates fell within the start and end sites of. However, multiple genes can overlap the same inferred intron for multiple reasons: either because Leafcutter introns are unstranded and genes from both strands can overlap them, or because of overlaps with smaller gene types such as miRNA, lncRNA, pseudogenes etc. which can reside within the coordinates of larger genes.

There were a total of 10,762 introns with significant sQTL associations which were overlapped by protein-coding or lncRNA gene coordinates, and some of these were overlapped by more than one of these classes of gene (*Table 4.3*). To choose the gene best matched to each intron, and to ascertain the extent of novelty of the events Leafcutter was able to identify, the inferred intron boundaries were compared against annotated exon coordinates from overlapping genes. 7,411 introns matched to known exon coordinates by both their start and end sites, 1,342 matched only by start site, 1,326 only by end site, and 683 were unmatched by either of their coordinates. These may represent novel, previously unannotated “cryptic” introns. 1,008 significant introns identified by Leafcutter were not fully-overlapped by any annotated gene from Ensembl gene build v88 mapped to GRCh38.p10. Of the 3,910 genes whose coordinates encapsulated an intron with a significantly associated sQTL, 3,781 were protein-coding and 129 lncRNA. Some genes harboured many different significant intron-level events (*Figure 4.10* a).

Leafcutter introns were also linked to a corresponding transcript in order to facilitate a comparison between Leafcutter sQTLs and the transcript-level sQTLseekeR events. Introns were assigned to any transcripts which overlapped them, then matches were further filtered by checking whether intron boundaries inferred by Leafcutter overlapped annotated exon boundaries for those transcripts. Slightly fewer significant introns (10,059) were able to be mapped to transcripts (3,720 from protein-coding genes and 129 from lncRNAs) using this methodology than were able

to be assigned to genes, leaving 2,771 which were not assigned to any known transcript. There were cases where an intron could potentially match multiple different transcripts (*Figure 4.10* b), and some transcripts contained more than one significant intron event (*Figure 4.10* c).

Gene type	Number of genes	Number of introns
protein-coding	4090	10373
antisense	306	728
lincRNA	163	433
transcribed unprocessed pseudogene	122	435
processed transcript	110	473
Other	88	298
NA	NA	1008
Total	4880	13748
Total unique	4880	12830

Table 4.3 Number of introns whose coordinates were overlapped by different classes of genes
The total number of introns in the table (13,748) is greater than the 12,830 unique significant introns because certain introns can be overlapped by multiple genes and therefore represented multiple times in the table.

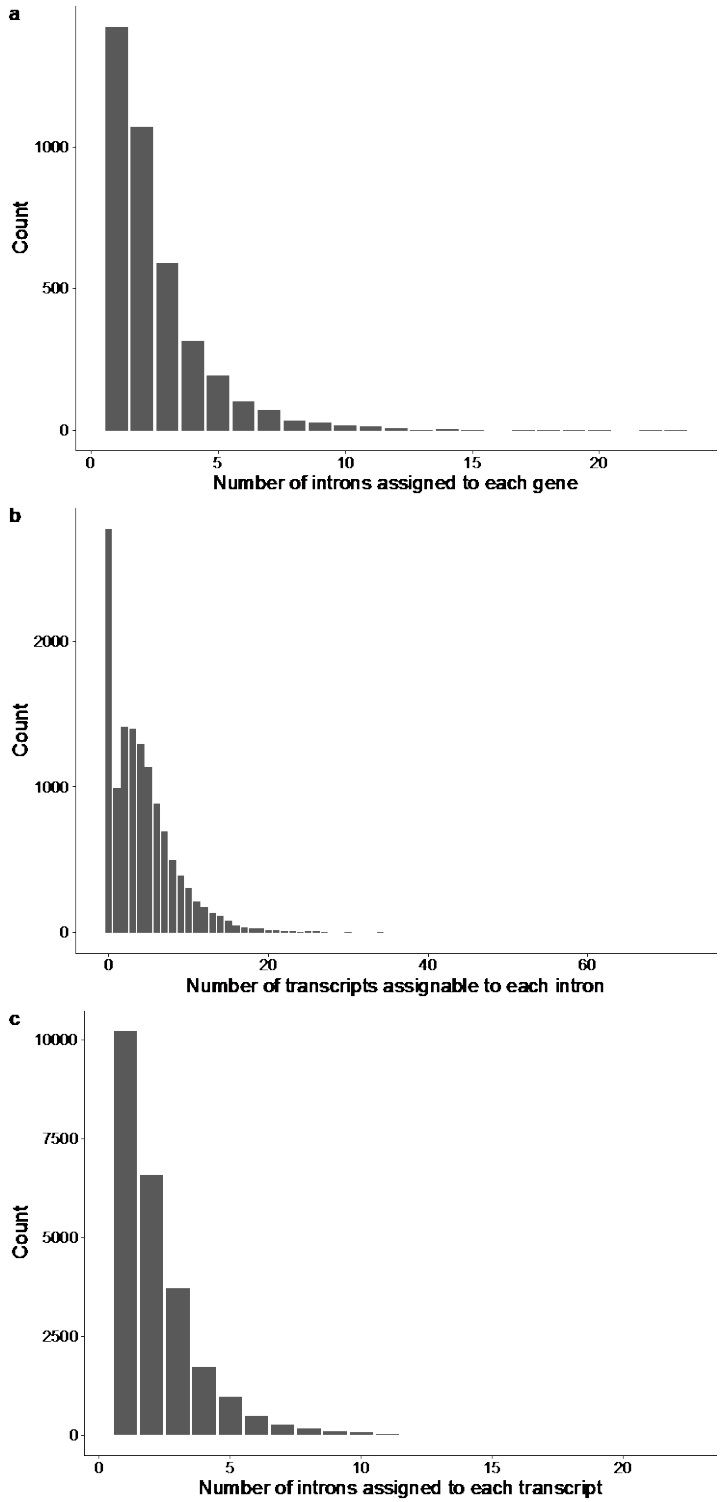


Figure 4.10 a) Number of FDR 0.05 significant Leafcutter intron events assigned to each gene based on agreement between intron coordinates and exon boundaries.
Mode 1.00, median 2.00 and mean 2.58 (n=3,910 genes)
b) Number of transcripts potentially assignable to each intron
Zero represents the 2,771 introns unable to be assigned to a known annotated transcript.
c) Number of introns assigned to each transcript
Median 2.00, mean 2.279.

4.3.5 Local distributions of events

SNPs associated with transcript-level sQTL events by sQTLseekeR were most commonly found within the body of genes (*Figure 4.11*). There were fewer events found up and downstream of genes, without any obvious drop-off in the rate of events throughout the 5kbp window.

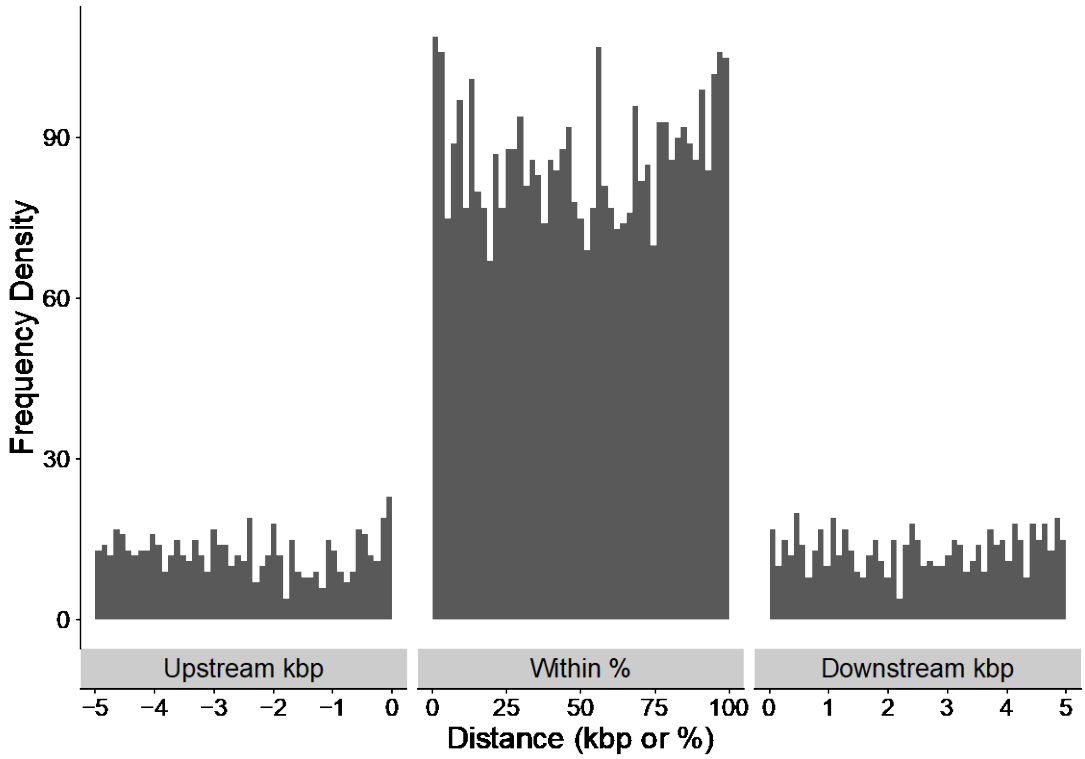


Figure 4.11 Distribution of sQTLseekeR events relative to gene body
Where sQTL SNPs fall upstream or downstream of the gene they relate to, the distances are plotted in kbp. Where a SNP falls within the gene body, the percentage distance along the length of the gene is shown. n=614 Upstream, n=4,231 Within, n=647 Downstream.

The search window set by FastQTL is fundamentally different to sQTLseekeR's, defined as a range around the start-site of each Leafcutter-inferred intron rather than the whole gene body plus a window. Because the length of introns can vary greatly, a larger window of ±100kbp is required with FastQTL to maximise the likelihood of achieving coverage over the gene-body of every gene. As a corollary, FastQTL searched much farther up or downstream of certain genes than sQTLseekeR, and a clear drop-off in the rate of events can be observed as distance up or downstream from a gene increases (*Figure 4.12*a). There was also a drop-off in significance of Leafcutter events as the distance from gene body increased (*Figure 4.12*b).

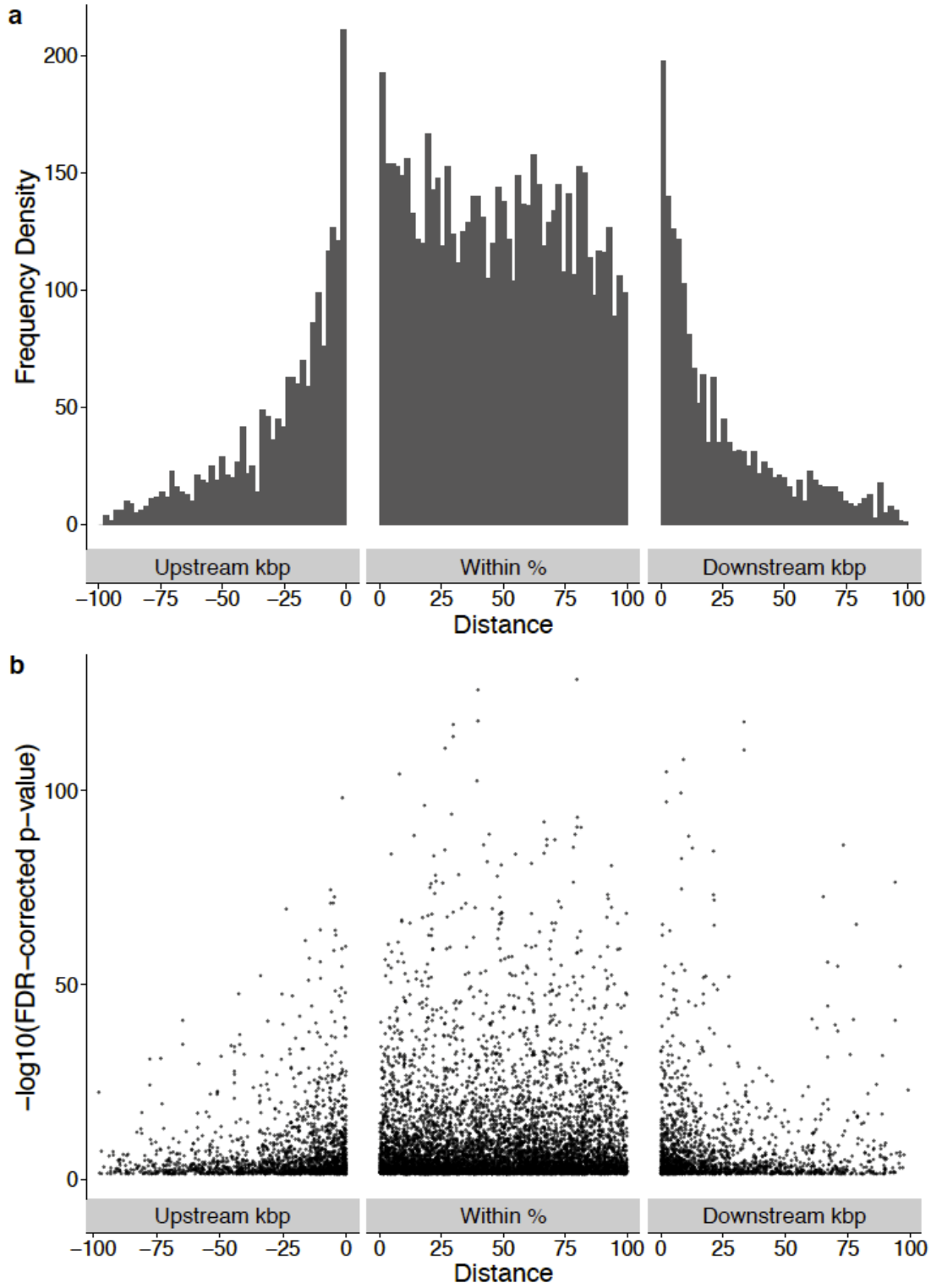


Figure 4.12 a) Distribution of Leafcutter events relative to gene body.
In cases of multiple potential mappings, the largest protein-coding gene was selected. Up and downstream distances between sQTLs and genes is plotted in kbp; sQTLs falling within gene bodies are plotted by percentage distance along the length of the gene. n=1,853 Upstream, n=6,475 Within, n=1,731 Downstream.
b) Significance of Leafcutter sQTLs against distance to gene body

In case the mapping from introns to genes was imperfect and masked any specific patterns in the data, the significance of events relative to simply the distance from the start-site of each intron was also plotted. This shows a clear peak around the beginning of the intron boundary, and a similar drop-off in significance as distance increases (*Figure 4.13*a). This plot is more indicative of the actual search window adopted by FastQTL, given that it only searches relative to the start site of each intron and is naïve to the relationship between introns and genes. It also appears that effect size of the Leafcutter events decreases further away from the intron start site (*Figure 4.13*b), which is to be expected given the relationship observed between these two metrics in *Figure 4.9*.

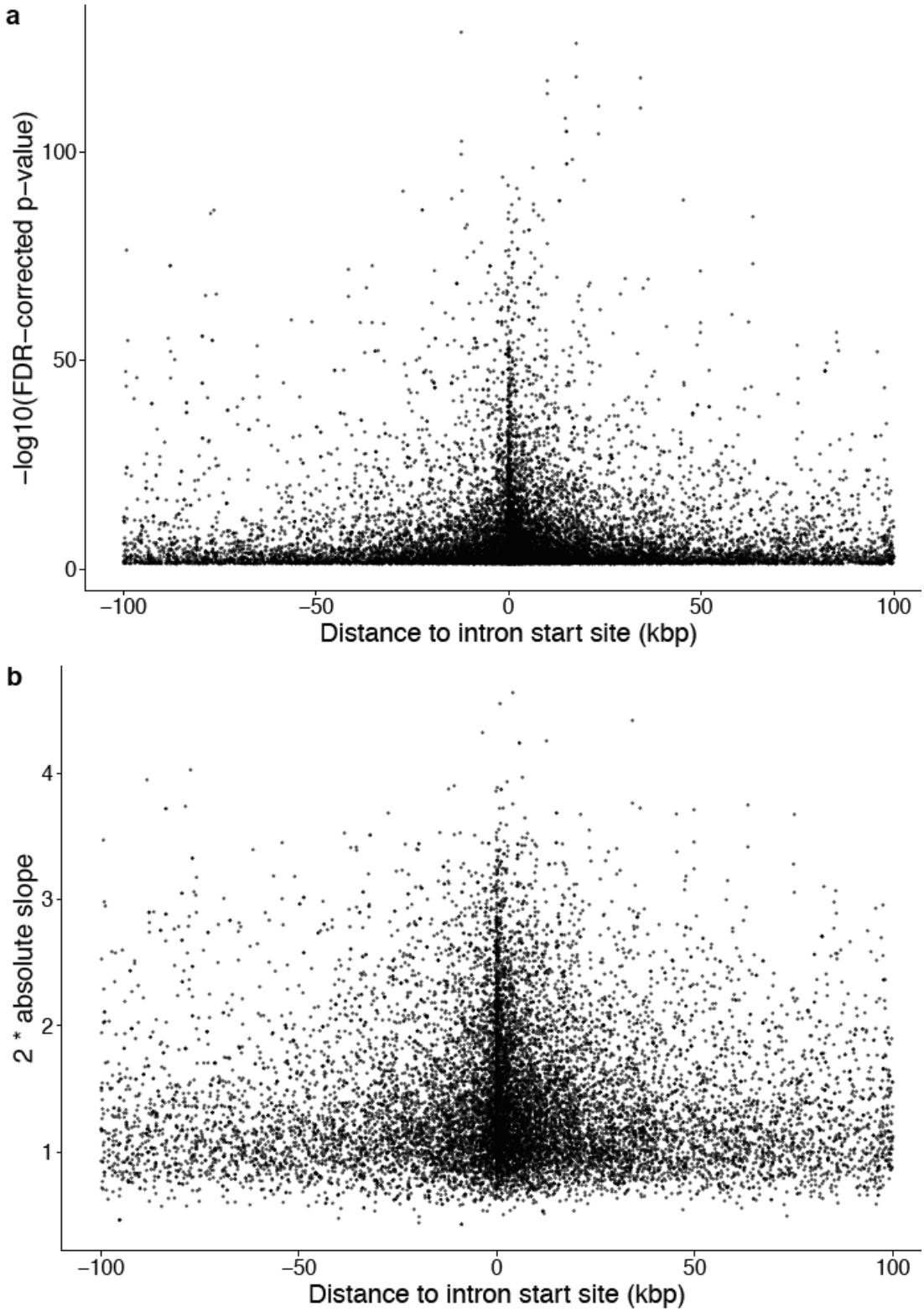


Figure 4.13 a) Significance of Leafcutter events relative to distance from intron start site
b) Effect size of Leafcutter events relative to distance from intron start site
The effect size is twice the absolute slope for the FastQTL linear association between genotype and Leafcutter intron usage ratio.

4.3.6 Genome wide distributions of sQTL SNPs

When viewed genome-wide, it is difficult to identify any hot-spots denoting clusters of significant sQTLs due to the issue of p-value banding from the sQTLseekeR algorithm, caused by there being a limit to the maximum significance which can be ascribed to an event using a non-parametric approach with 221 samples (*Figure 4.14*a).

The Leafcutter results do not suffer from p-value banding, and so hot-spots of sQTLs are more clearly discernible (*Figure 4.14*b). The MHC region on 6p21 is the densest region of sQTLs, however it does not contain the most highly significant events.

There were fewer highly significant events identified by either package on chromosome 23 (X). This could be because of the sparsity of variants probed for by SNP arrays for the X chromosome, or because with ~50% of the samples being male, there is decreased power on the X chromosome to call sQTLs because there are fewer heterozygous individuals. Overall expression will also be lower on the X chromosome because males will have no expression from female-specific X chromosome genes, meaning that fewer genes will pass the thresholds set by the packages for inclusion in the sQTL-detection algorithms.

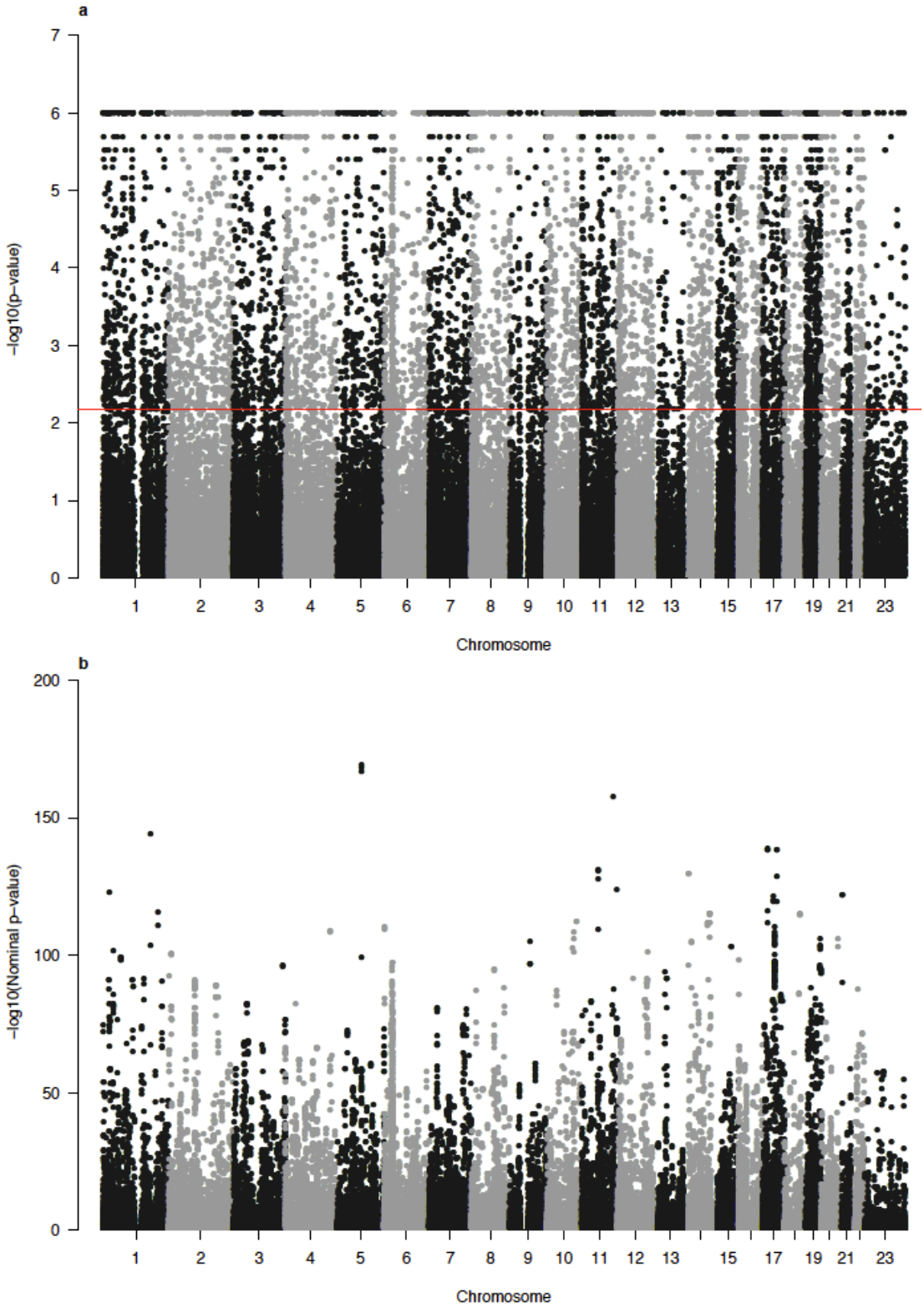


Figure 4.14 Manhattan plots of sQTL distributions
FDR 0.05 significant results and a random selection of 5% of the nominally significant events for **a)** sQTLseekR (red line = genome-wide significance threshold) and **b)** Leafcutter (a single genome-wide significance threshold is not applicable due to the FDR correction being tailored to each individual intron). Plots generated with the qqman package[377].

Despite the Manhattan plots being of differing quality between the two sets of results, when examining the distribution of the total number of events identified per chromosome, a strikingly similar pattern is observed (*Figure 4.15*). It is clear that the potential for finding sQTLs is dependent on the number of protein-coding genes available to be tested.

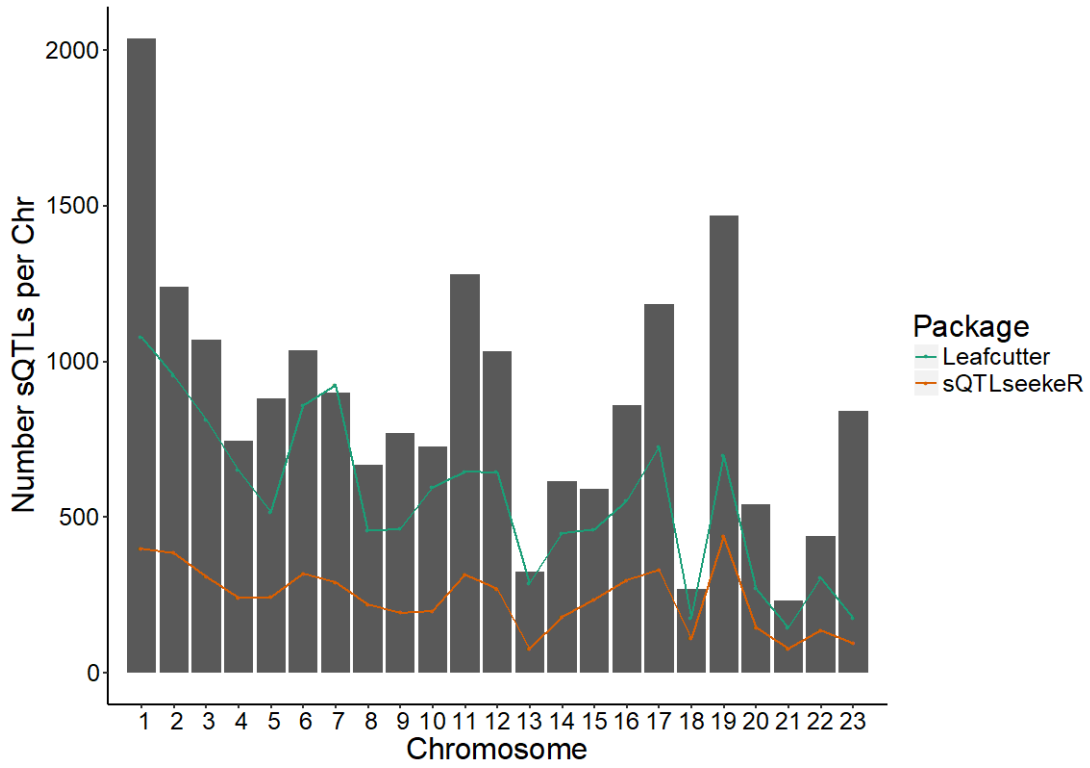


Figure 4.15 Number of sQTL events found by each package per chromosome
Background bars denote number of protein-coding genes located on each chromosome. The number of sQTLs found by Leafcutter on Chromosome 7 is greater than the number of total protein-coding genes on that chromosome because Leafcutter identifies intron-level sQTL events independent of gene annotation. n=5,492 sQTLseekeR events, n=12,830 Leafcutter events, n=19,741 protein-coding genes.

4.3.7 Comparisons between sQTLseekeR and Leafcutter sQTLs

There was an intersect of 1,387 genes for which sQTLs were identified by both packages, which is equivalent to 40.5% of the sQTLseekeR genes and 35.5% of the Leafcutter genes (*Figure 4.16*). Of these, sQTLseekeR identified sQTLs in 209 lncRNA genes, whilst Leafcutter identified 129. The intersect of lncRNAs was 39, which is 18.7% of the sQTLseekeR lncRNA genes and 30.2% of the Leafcutter (*Figure 4.16*).

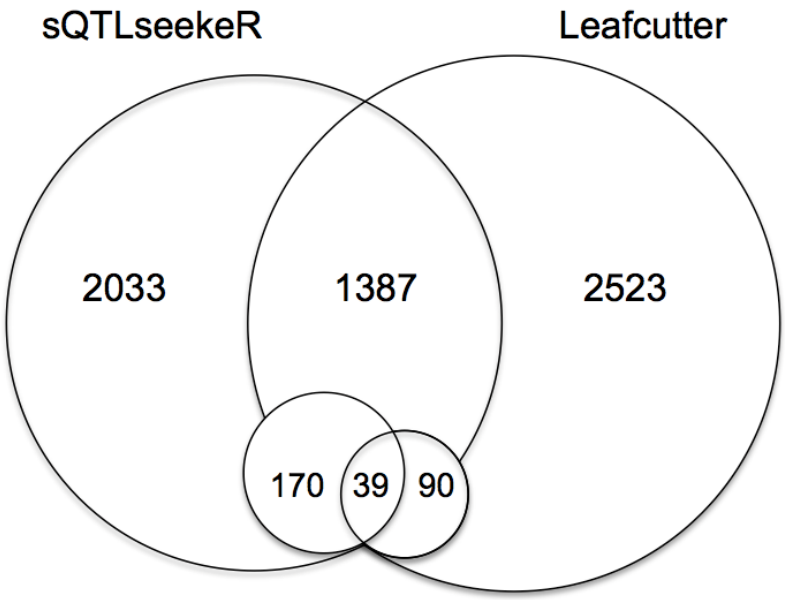


Figure 4.16 Intersect of genes with associated sQTL events identified by sQTLseekeR and Leafcutter.
Protein-coding genes are in upper circles, and lncRNA in lower.

The degree of correspondence between sQTL events identified by the two packages was compared by correlating the effect sizes of sQTLs where the intron identified by Leafcutter matched at least one exon boundary of one of the transcripts involved in an sQTLseekeR event (*Figure 4.17*). This produced a Spearman Rho correlation of 0.087 (n=8,144, p=3.61e^{-15}). The low correlation could be explained by multiple factors. The MD values of sQTLseekeR are scaled to a ratio between 0.0 and 1.0, whereas the effect sizes for Leafcutter are based on the slope of the linear correlation calculated by FastQTL in normalised space, so it is not surprising that the two are poorly comparable. Also, because Leafcutter introns have the potential to correspond to multiple transcripts and therefore to multiple different transcript-level sQTL events, there is rarely a perfect one-to-one comparison between effect sizes for the exact same change in alternative splicing of a given feature (transcript or intron). One single sQTLseekeR MD value for a complex alternative splicing event may be represented across multiple Leafcutter sQTLs.

If correspondences are filtered to only include the most significant of the potentially corresponding pairs of events between packages, then the correlation increases to 0.114 (n=2,337, p= 3.56e^{-08}). The correlation also increases further to 0.216 if only events with exactly the same “lead” associated SNP are used (n=123, p=0.0163). However, the correlation still remains low, and this serves to further highlight the

differences between the outputs of the two algorithms and the challenges in drawing relevant comparisons between them.

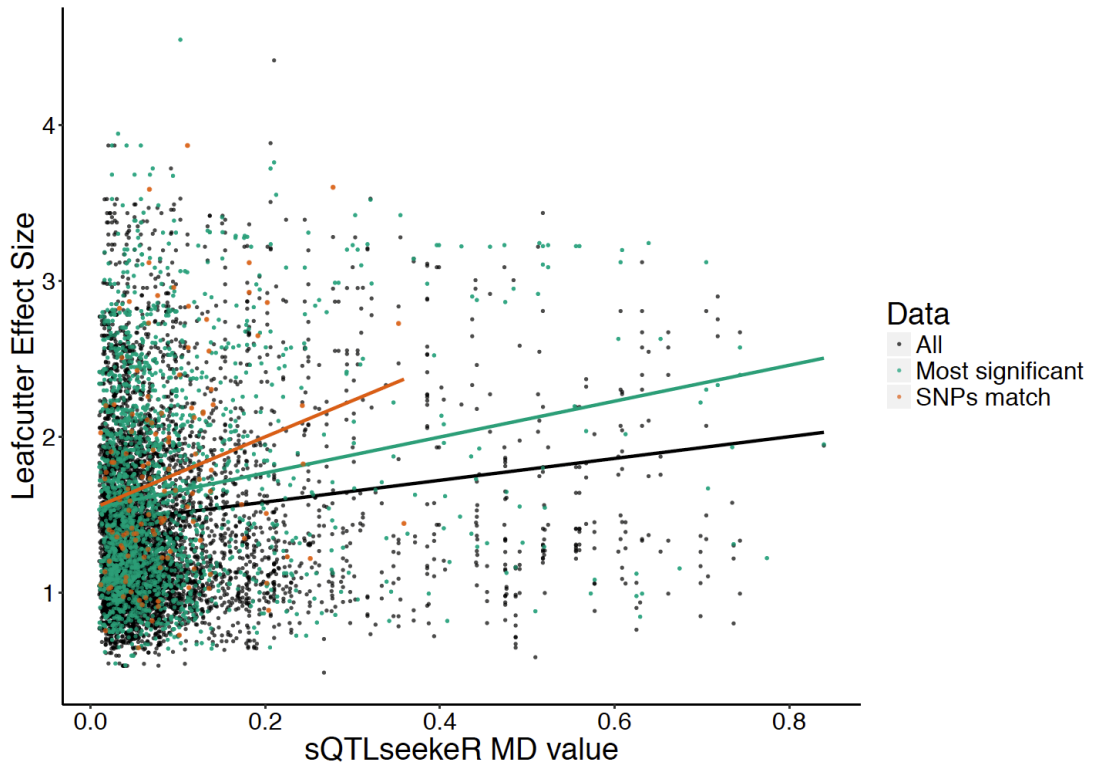


Figure 4.17 Effect size correlation of events between packages
Best fit lines are a linear model fitted by the "lm" linear model function in ggplot2.

4.3.8 Filtering of sQTL events

Filtering of sQTL events was carried out so as to supply a higher-confidence list of sQTLs for the functional characterisations carried out in the following chapter.

sQTLs derived from lowly expressed genes were excluded. There was a negative correlation between effect size and mean gene expression level for events identified by sQTLseekeR (*Figure 4.18*). This is likely because events supported by only a low number of counts have less scope to generate smaller ratios of transcript expression for a given genotype group, and therefore the MD effect size can only adopt a limited range of values. Conversely, there is a positive correlation between effect size and the number of read counts supporting introns associated with sQTLs by Leafcutter + FastQTL. The tool's effect sizes are not ratios, but are calculated from linear associations; therefore the more reads available for a given intron, the steeper the gradient of linear association that is can be produced (*Figure 4.19*).

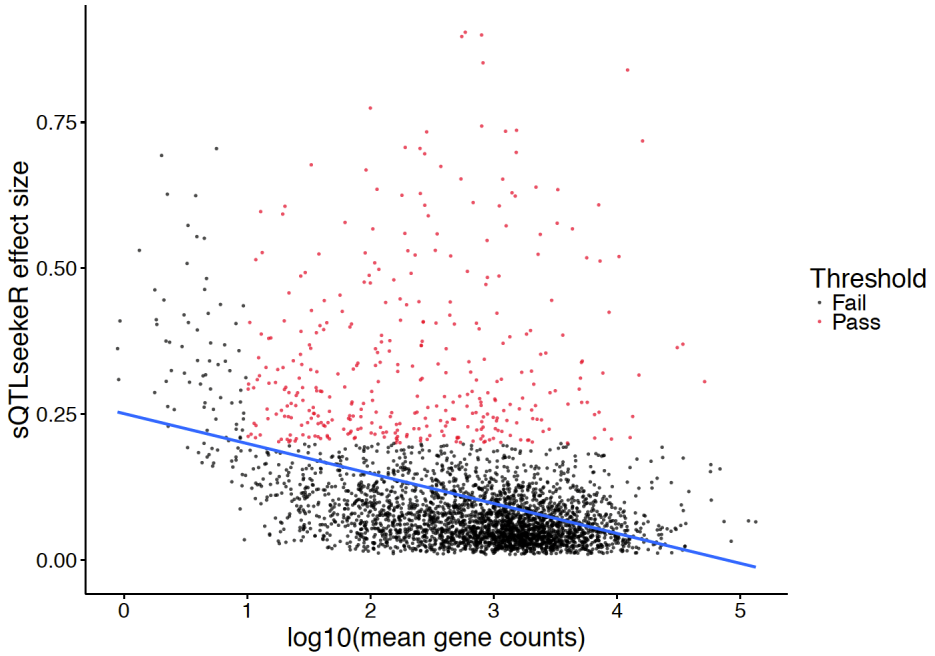


Figure 4.18 sQTLseekeR effect size against mean gene expression across 221 samples
Salmon transcript-level counts were aggregated to gene-level by tximport and the mean gene expression calculated across 221 samples. Blue line represents a linear model fitted by the "lm" function in ggplot2 with Spearman Rho: −0.369, p−value: 1.44e−110. Red points passed a threshold of 0.2 effect size and 1.0 $\log_{10}$(mean gene count).

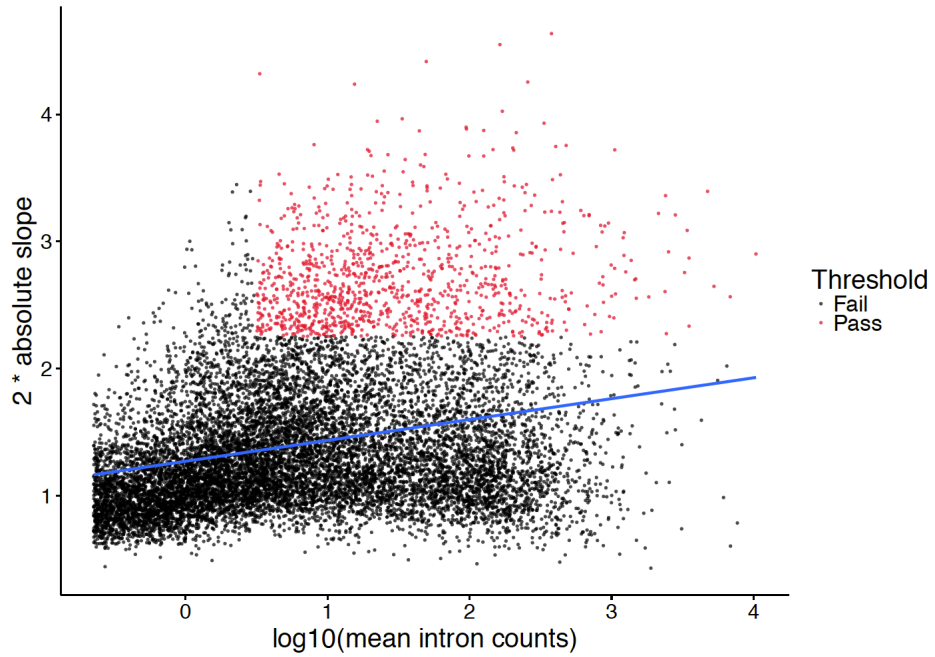


Figure 4.19 Leafcutter effect size against mean read counts supporting intron excision
Mean read counts assigned to introns by Leafcutter across the 221 samples are plotted against 2* the absolute slope of correlation. Blue line represents a linear model fitted by the "lm" function in ggplot2 with Spearman Rho: 0.289, p−value: 4.01e−246. Red points passed a threshold of 2.25284 effect size (2*absolute slope) and 0.5 $\log_{10}$(mean intron count).

4.4 Discussion

4.4.1 Reliability of sQTL identification

Accurate quantification of transcript expression is key to the identification of sQTLs. The Leafcutter authors benchmarked their algorithm using either OLEGO or a two-pass STAR alignment and found little difference between the two: 4.8% of junctions differed by a fold difference of 1.1 and only 0.94% of junction counts differed by a fold difference of >2.0[260]. This study only used a single-pass STAR alignment[9], so this may have slightly reduced the number of reads mapped and so the scope of Leafcutter to infer sQTLs relating to the usage of novel intron junctions; however OLEGO also only runs a single-pass alignment, so any gains from the additional alignment pass would likely have been negligible.

There is some intrinsic uncertainty in the calling of variants from fluorescence-based SNP arrays and in the imputation of genome-wide variants from exome-only chips. The simple conversion of decimal genotype dosages to integers may have masked some uncertainty in the calling of variants; however it was a necessary step for input of the data into sQTLseekeR. If a threshold had been set requiring the SNPs to have dosages ≤0.1 away from an integer for >90% of the samples in order to be accepted, then this would only have removed 5.17% of the total variants. Thus the scope for false positive observations as a result of not having imposed any such thresholds is limited. FastQTL was run directly on the original decimal genotype dosages without any filtering, and so had the opportunity to account for the uncertainty of decimal values in its linear model.

Allele-specific expression (or biases in detection or half-life of certain allele-specific transcripts) may influence the identification of sQTLs and introduce false positive results. The authors of Leafcutter suggest running a tool called “WASP” prior to any expression-dependent QTL analyses in order to identify and screen out such variants[378], but they did not apply WASP for their 2018 analysis of GEAUVADIS CEU expression data to maintain comparability with previous analyses[260]. Panousis *et al.* demonstrated in their 2014 study that allelic bias arising from preferential mapping of reference alleles over alternative alleles is not a major source of false positive eQTLs, with only 3.67% fewer genes identified as having an eQTL with

[9] STAR alignment performed by Dr Alison Meynert.

allele-specific reads filtered out[379]. Based on this observation of minimal gains, correction for allele-specific expression was not performed in this study - although it could be argued that mapping biases of allele-specific reads may have more potential to influence sQTLs than eQTLs because the sequences of individual transcripts matter more when not being summed to total gene expression.

There were often multiple variants associated with each alternative splicing event (*Figure 4.5* and *Figure 4.10*), a common pattern in QTL studies[200]. In this project, where a "lead" SNP was chosen, the significance and effect size were used to prioritise the most likely variant. Lappalainen *et al.* showed in their 2013 analysis of QTLs in GEAUVADIS that using these metrics did identify the most likely causal variant in between 34% and 74% of cases, depending on the population analysed, although they derived genotypes from whole genome sequencing data, and therefore had access to more variants[254]. They report that the lead SNP was present in WGS from 1000 Genomes but not on an Omni 2.5M SNP array in 81% of cases. Variants in this study were imputed to 1000 Genomes to increase the potential to identify causal sQTL SNPs not captured by arrays alone. The instances in this analysis where multiple SNPs were associated with an sQTL event with exactly the same significance and effect size likely means the variants were in particularly high LD with each other, and there were the same numbers of the 221 individuals falling in each of the genotype groups for the linked SNPs, and so their effects were not able to be differentiated. Other more advanced methods do exist to fine-map the causal SNP for a QTL event, for example RASQUAL uses other genomic features such as DNA-accessibility in the form of ATAC-seq to narrow down the most likely candidate causative variants for eQTLs[380]. A similar approach could have been adapted to sQTLs, however high depth ATAC-seq was not available from normal colonic mucosa at the time of this analysis.

De novo transcriptome assembly with tools such as Trinity[381] or StringTie[382] was not performed for this study because the RNA-seq data was obtained from normal mucosa samples, and the transcriptomes were therefore assumed to be normal. If matched tumour RNA-seq were used for future comparisons with this normal data, then there would be greater motivation to search for novel splicing patterns. Some authors suggest that even when analysing normal tissue there is novelty to be found: when analysing non-tumour RNA-seq from 75 individuals from 1000 Genomes consortium, Stein *et al.* were able to identify 437 novel splice junctions

that were not mappable to the human reference genome by constructing “personal genomes”[383]. This implies some additional novelty may have been missed by not taking this approach in this study, however it is likely to be small.

The sQTLseekeR authors propose that their algorithm should be reasonably immune to batch effects because it relies on transcript expression ratios which are internally controlled against the total expression of each gene. When identifying sQTLs from Leafcutter-inferred intron usages, FastQTL is able to use covariates to account for potential confounding factors, of which the clearest component from PCA corresponded to batch (*Figure 4.4*). Therefore, there is an argument to be made that each batch could have been analysed separately so as to avoid this potential confounding factor, and then the sQTLs from each combined together. However, both batches were combined so as to increase power when testing for differences in transcript expression or intron usage ratios between genotype groups, and to have higher likelihood of being able to analyse rarer variants. When calling sQTLs from GEAUVADIS LCLs in their 2014 study, the sQTLseekeR authors ran their analysis on each of the subpopulations of CEU, FIN, GBR, TSI and YRI, which consisted of 91, 95, 94, 93 and 89 samples respectively. They found a 92% sharing of sQTLs between groups[247], which in addition to demonstrating the transcriptional similarities of the populations also implies that power would be wasted analysing smaller cohorts separately and finding the same sQTLs repeatedly when combining the samples may have allowed other, rarer sQTLs to be identified. Similarly, when deriving QTLs from 44 tissues, the GTEx Consortium found that greater numbers of eQTLs were discovered when expression data from multiple tissues were combined together than when each tissue was analysed separately[200]. In a preliminary analysis carried out at the beginning of this project, sQTLseekeR was able to identify sQTLs for 1,658 unique genes from 92 samples from batch 2013152 at FDR 0.05. This is roughly half the 3,420 genes with sQTLs identified from the combination of batches 2013152 and 10525, demonstrating the utility of combining the two cohorts together for increased power.

Instead of relying on transcript expression quantification at all, methods have been derived to predict alternative splicing based on sequence alone. Jaganathan *et al.* trained a 32-layered deep neural network to predict patterns of alternative splicing from pre-mRNA sequences, using 5kbp-worth of sequence up and downstream of each base to predict whether it constituted a splice donor or splice acceptor. The

events it predicted had a 75% validation rate from RNA-seq[216], and when perturbing the sequences artificially, it was found that the bases most contributing to the model's predictions were upstream splice acceptor sites, or the binding sites of SR-family proteins or branchpoint spliceosome binding motifs[216].

4.4.2 sQTLseekeR events

In their 2014 analysis of GEAUVADIS LCLs, the sQTLseekeR authors identified an average of 2,900 sQTLs at 5% FDR for each of the populations they analysed[247]. This is likely fewer than the 5,492 different significant transcript-pair sQTLs identified in this study because they analysed smaller individual populations separately. They also had a lower number of genotyped SNPs (1.3M as opposed to 3.9M), and had shorter read lengths and a lower depth of expression data: 75bp paired-end RNA-seq with a mean of 48.9M reads per sample compared to 100bp or 150bp paired end reads with means of ~130M and ~150M mapped reads per sample (batches 2013152 and 10525 respectively). Increased read length has been demonstrated to improve splice-junction detection by Chhangawala *et al.* based on analysis of ENCODE samples, with 100bp reads shown to identify up to 25% more splice junctions compared to 75bp[384]. There were also differences in the alignment and quantification of transcript expression: they used the GEM[385] aligner and quantified transcript expression from the genomically-aligned reads with Flux Capacitor[249], as opposed to the alignment-free quantification via Salmon[301] which was used in this study.

There was no clear drop in the rate of sQTLs identified along the 5kbp search window up or downstream of genes for sQTLseekeR (*Figure 4.11*). Searching within sequences proximal to gene bodies was justified because this is where most variants influencing regulation are expected to occur[254,386], however the lack of drop in rate could imply that it may have been useful to extend the search region further - though perhaps not as far as the 100kbp window that Leafcutter used, which displayed a clear drop-off in events (*Figure 4.12*). Statistical power may have been diminished by FastQTL evaluating more SNPs farther away from genes and consequently requiring greater multiple testing correction.

DRIMSeq is a newer algorithm than sQTLseekeR, which instead of using a non-parametric MANOVA to analyse expression ratios of each transcript, directly models the transcript expression of all isoforms of a gene using a Dirichlet Multinomial

distribution[246]. An advantage it confers over sQTLseekeR is that the model can include a representation of the expression of each transcript and account for this when quantifying uncertainty in the estimated expression ratio which the transcript contributes to total gene expression. Expression ratios calculated by sQTLseekeR from lower-expressed transcripts will have lower certainty and wider intervals of confidence, which is not currently taken into consideration by the algorithm. However, the setting of low expression thresholds and post-hoc filtering should have reduced the contribution of inaccurate ratio quantification to this study. The risks of inaccuracy are also greater when there are fewer samples, which further justifies the combination of both batches for the purposes of this analysis. When analysing RNA-seq from 91 CEU samples and 89 YRI samples, DRIMSeq found fewer genes containing sQTLs than sQTLseekeR. sQTLseekeR was able to find more associations in lower expressed genes and in genes with fewer different expressed transcripts, and found more associations with SNPs farther away from the nearest proximal exon[246]. It was desired in this study to perform as comprehensive as possible a survey of alternative splicing events, hence the utilisation of sQTLseekeR is justified.

One clear limitation of the sQTLseekeR package is its lack of ability to accept covariates, though other transcript-level sQTL association tools such as DRIMSeq do not include covariates either[246]. It may not be an essential addition, however. Singh *et al.* studied eQTLs in colonic tissues using linear regression of microarray probe intensities against genotype based on an additive model with either no covariates, or covariates including age, gender, population-level principal components and clinical metadata including disease status (23 patients had ulcerative colitis, 16 Crohn's disease, 33 no disease)[222]. They concluded that there was a reduction of power when all possible covariates were included, and that there was robust concordance of results between models with or without the inclusion of covariates, and so they decided to report as their final results the direct pairwise regressions between array probe intensities and variants[222]. This lessens the motivation for adding covariates to sQTL detection algorithms which do not currently account for them, particularly sQTLseekeR, which would not be able to accept them in its current non-parametric form.

4.4.3 Leafcutter events

In this study, Leafcutter inferred the presence of 175,792 different introns exhibiting differential excision. “SplAdder” is a similar programme which also infers differential intron useage based on RNA-seq alignments, however it first employs a genome annotation file to building a splicing graph, then uses genomically-aligned RNA-seq reads to assign confidence to predicted introns and prune or augment the graph appropriately[387]. SplAdder was used by Lehmann *et al.* to search for cancer-specific splicing patterns by comparing the frequency of intron usage from 282 kidney renal clear cell carcinoma (KIRC) samples to normal samples from the same patients, plus 140 GEAUVADIS and 460 ENCODE RNA-seq samples. 160,208 introns with alternative usage were identified - similar to the 175,792 seen in this study using Leafcutter - though with less stringent thresholds of only requiring an intron to be supported by at least 10 reads in any one of the 882 samples[248]. They used an older version of the genome build as a basis for SplAdder to augment (Gencode version 14, which corresponds to the Ensembl release version 69 from the 4th quarter 2012), and each of the TCGA samples had fewer reads than this study with approximately 70M.

Novelty of introns inferred by Leafcutter

The Leafcutter authors found in their 2018 analysis of GTEx tissues that 10.8%, 19.3% and 48.5% of the introns they inferred from pancreas, spleen and testes (a notable outlier) respectively were previously unannotated in either of GENCODE v19, Ensembl v75 or UCSC v19 gene builds. Leafcutter was also used by Raj *et al.* to investigate alternative splicing in brain autopsy samples from 450 individuals partaking in prospective cohort studies of aging, and 30% of the 53,251 intron clusters identified were novel to the extent that they had not been previously reported in other sQTL studies[265].

Using *vast-tools*[388], an algorithm which similarly to Leafcutter is designed to predict alternative usage of exonic sequences from genomically-aligned RNA-seq reads, the creators of the Vertebrate Alternative Splicing and Transcription Database (VastDB) found that 13.9% of the alternative-splicing events they identified from 108 human individuals were not previously annotated in Ensembl[389]. In this study, 7.86% of Leafcutter-inferred introns with a significant sQTL did not overlap any gene annotation from Ensembl version 88, and so would be considered novel. Of the 10,762 introns which fell within the coordinates of protein-coding or lncRNA genes,

68.9% shared both their intron boundaries with exon coordinates annotated in Ensembl GRCh38v88. 24.8% shared at least one boundary with an exon coordinate and 6.35% shared none. Taking the sum of these last two categories would make 31.2% of introns which could be considered at least partially novel. This falls between the extremes identified by the Leafcutter authors of 10.8% to 48.5% from pancreas and testes. The numbers in this study may be higher than they observed for other non-germline tissues because it was not assessed whether introns identified by Leafcutter shared intron boundaries with annotated genetic elements outside of the coding sequence such as alternative 3' or 5' UTR regions. Some intron sQTL events may have occurred due to changes in usage of known untranslated elements and therefore estimates of novelty here may to some degree be inflated.

Numbers of sQTLs identified

In their 2016 study the Leafcutter authors identified 2,893 sQTLs in 2,313 genes at 10% FDR from the 89 LCL samples from the GEAUVADIS YRI population[261]. In their 2018 study, the same group identified 5,774 sQTLs at 5% FDR from 372 LCL samples from the GEAUVADIS EUR population[260]. This is compared to 12,830 from this analysis of 221 samples. As discussed in relation to sQTLseekeR, the GEAUVADIS data has deficits when compared to this study in terms of SNP density and read depth, which perhaps limited the number of associations they were able to find. Prior to running the sQTL associations in their 2018 study, Leafcutter had identified 42,716 intron clusters with alternatively excised introns. This is only marginally fewer than the 47,977 clusters able to be identified by this study, implying that perhaps the SNP density rather than the numbers of aligned reads was more of a limiting factor in identifying sQTL events, because close to the same number of intron clusters with alternative patterns of splicing was identified.

Lehmann *et al.* only identified 915 splice QTLs from KIRC and normal samples from 282 individuals[248], using PSI of exons as the phenotypic measure. However they only calculated associations with 458,266 variants, called from exome-sequencing by the HaplotypeCaller[390] tool from GATK[391] for germline, and the MuTect[392] tool for somatic variants. This is further suggestive that the number of variants available to be associated with splicing events is a limiting factor in sQTL discovery.

Using Leafcutter and FastQTL Raj *et al.* identified 9,028 sQTLs in 3,006 genes from brain samples of 450 individuals at 5% FDR[265]. This corroborates the result in this study that multiple significant sQTL events can emanate from the same intron cluster from the same genomic coordinates (*Figure 4.8*). Their analysis may have found fewer significant sQTLs than the 12,830 from this study because even though they sampled more individuals, their samples came from autopsy as opposed to during live surgery. The negative correlation between RNA-integrity and post-moretem interval has been well documented across multiple tissues by the GTEx Consortium[393], and in colonic tissues specifically by Musella *et al.*[394], meaning RNA may have been more degraded and so increasing the challenge of extracting signals of alternative splicing.

4.4.4 Comparison of sQTLseekeR and Leafcutter

Although the two packages have been run by both of their respective authors on GEAUVADIS LCL expression data, sQTLseekeR was run individually for each of the 5 populations whilst all 4 European populations were combined for the Leafcutter 2018 study. Both have been run individually on the 89 samples of the YRI GEAUVADIS LCLs, however the reported results were at 5% FDR from sQTLseekeR and 10% FDR from Leafcutter[247,261]. This project offers the opportunity to compare the performance of these two packages on the same sample of 221 individuals at the same level of FDR.

Leafcutter identified more significant sQTL events than sQTLseekeR at FDR 5%: 12,830 as opposed to 5,492. This could be due to multiple factors, such as its larger search window, or its ability to identify novel intron events whilst sQTLseekeR is limited to finding events relating to known annotated transcripts. Also, there may be multiple different intron-level Leafcutter events which are captured by just a single sQTLseekeR event because it is able to identify more complex splicing changes than Leafcutter. The majority of events identified by sQTLseekeR were classified as being “complex” (*Figure 4.7*), and would therefore have required multiple Leafcutter events to capture. The Leafcutter authors noted in their 2016 study that 60% of the sQTLs they identified related to simple exon skipping of one or more exons in series, 20% of events corresponded to alternative splice acceptor or donor sites for a single exon, 10% corresponded to an alternative exon at the 5’ or 3’ end of the gene, and only the remaining 10% could be classified as a more “complex” event[261]. When using Altrans and sQTLseekeR, the GTEx Consortium also observed the

number of sQTLs identified from the exon-based algorithm to far exceed those from the transcript-based (1,900 vs 250)[200].

The agreement of 40.5% between genes for which sQTLs were identified between the two packages seems reasonable given the substantial differences in the way their algorithms operate, and the fact that Leafcutter was technically designed to run independently of a gene build meaning that genes could only be assigned to events post-hoc (Section 4.3.4). sQTLseekeR used raw read counts as quantified by the pseudo-aligner Salmon and did not perform any normalisation before using a non-parametric significance test without any associated covariates. In contrast, Leafcutter used genomically aligned reads to infer intron excision ratios, which were then centred and normalised prior to running through a parametric linear model with associated covariates. The previous chapter explored the lack of perfect correlation between genomically-aligned and alignment-free expression quantification when correlating Cufflinks vs Salmon, which may further explain differences in the sQTL events identified. The GTEx consortium observed a similar agreement to that seen in this study of 36% of sQTLs found by both sQTLseekeR and Altrans, though this was only for exon-skipping events, which both packages should be proficient at identifying[200]. A group investigating the effects of toxic lead doses on the development of *Drosophila melanogaster* used both whole-transcript-based and PSI-based methods to identify sQTLs. They found a similar concordance between the two methods, identifying 374 via transcripts and 974 via exons, with an agreement of 112 (29.9% or 11.5% respectively)[395].

sQTLseekeR was more adept at finding sQTLs associated with lncRNAs, identifying 209 compared to 129 by Leafcutter+FastQTL. This could be expected given the trend for lncRNAs to have fewer introns than regular protein-coding genes, therefore there would be fewer opportunities for Leafcutter to call an sQTL based on intron excision (mean number of exons according to Ensembl GRCh38v88: protein-coding: 27.8, lncRNA: 4.45[328]). The 39 lncRNAs which both packages identified may warrant further follow up, as a recent study by Huyghe *et al.* linked rare variants in lncRNAs to colorectal cancer risk, including one at 7p13 which has been demonstrated to promote pancreatic cancer proliferation through epigenetic repression of KLF2[231]. The potential of lncRNAs to influence the initiation and progression of colorectal cancer is also highlighted by a recent paper from Forrest *et al.,* which identified the long non-coding RNA *lincDUSP* as being over-expressed from an analysis of 22

CRC tumours vs 22 normal mucosa controls[396]. With *lincDUSP* knocked down, patient-derived cancer cell lines showed reduced cell viability and increased susceptibility to cell death and apoptosis[396].

4.4.5 Effect size of sQTLs

The effect sizes of the majority of sQTLs identified by both packages were modest. The sQTLseekeR authors arbitrarily ascribe any event with an MD value of ≥0.2 as being probably biologically relevant. 9.27% of the sQTLseekeR events in this study had an MD value of ≥0.20, 14.8% ≥0.15 and 26.2% ≥0.10 (*Figure 4.6*).

The effect sizes correlated poorly between the packages (*Figure 4.17*), likely due to the different scales and distributions of effect size calculated by each package. Being a non-parametric test based on ranks, there was some vertical banding observed in the sQTLseekeR effect sizes, whereas having made use of a linear association, the effect sizes for Leafcutter's results were continuous. Perhaps the effect sizes may have correlated better if they had been transformed to a more comparable scale to one another, e.g via quantile normalisation. Additionally, the fact that there was never a perfect 1-to-1 relationship when trying to link events identified by the packages likely contributed to the poor concordance.

In the supplementary methods of their 2016 paper, the Leafcutter authors define effect size of their sQTL events as being twice the slope of the linear regression against genotype, as calculated by FastQTL[261]. They found that only 13.8% of their sQTLs had an effect size ≥10%, whilst 76.6% were over 1%, meaning that 23.4% of the events they analysed had an effect size of ≤1%. They accepted that events with an effect size ≤1% are "expected to have modest impact in general", and justified their inclusion in their analyses by claiming it is likely that certain sQTLs might have larger effects at different developmental stages or in other cell types, or that certain changes in splicing could result in the generation of toxic pepties which would have a large effect even at low levels. This study is specifically concerned with sQTLs presented in the colonic mucosa as it is the site of initiation of colorectal cancer, therefore low-effect size sQTLs in this tissue are of considerably less relevance to this study. Low effect size sQTLs are liable to be dominated by simple fluctuations of noise in the expression ratios of transcripts or introns (*Figure 4.18* and *Figure 4.19*), hence thresholding of low effect size events is important prior to functional characterisation.

4.4.6 Thresholding

Thresholds were applied to sQTLs for two reasons: firstly to address expression thresholds which were potentially too lenient during intial sQTL identification, and secondly in an attempt to filter for only the most functionally relevant events. The distributions of effect size against expression for both packages display clouds of points which could feasibly have been driven by noisy associations of low-effect size changes in transcript expression or intron utilisation (*Figure 4.18* and *Figure 4.19*).

It is not desirable to set arbitrary thresholds, however the approach taken was considered favourable compared to potentially reaching misleading conclusions influenced by false-positives. An alternative strategy to find naïve thresholds may have been to fit an unsupervised Gaussian mixture-model to the effect sizes, and if there was a divergence between two components, this would have indicated a suitable division between two separate populations of low and high effect size events. However, exploratory plots of these distributions indicated that a clear divergence was unlikely.

A further strategy may have been to identify a threshold at which the majority of sQTL events displayed “stable” effect sizes. It would be predicted that sQTLs with larger effect sizes would tend to always have the same transcript most highly expressed in relation to a particular allele, whereas conversely it could be envisaged that sQTLs with lower effect-sizes values (e.g. an MD value around 0.01) may simply be due to noise, and therefore could fluctuate which transcript is most highly expressed if there was repeat sampling of the same individual. If individuals were sampled at multiple timepoints, then it could be tested what effect size threshold produces sQTL events which are most “stable” over time. Such a test would require the same individuals to be repeatedly sampled at multiple timepoints with no external treatments applied. 50 individuals from batch 10525 were sampled at 0, 6 and 12 week timepoints; however they had received interventional vitamin D supplementation as part of the SCOVIDS study which could confound any conclusions, and additionally the relatively small number of individuals is not ideal for accurate quantification of sQTL events.

Other sQTL studies rarely report applying filters to their events, or plot distributions showing the effect sizes of identified sQTLs. Some authors openly acknowledge the propensity towards low effect size events in their results, though without imposing

any thresholds their conclusions risk being spurious or unreliable. When identifying sQTLs from normal kidney tissue and kidney renal cell carcinoma, Lehmann *et al.* identified 915 sQTLs, of which they concede only 251 had an effect size >5%[248]. As discussed, the Leafcutter authors comment in the supplementary methods of their 2016 paper that 86.2% of their sQTLs have an effect size of <10%, and 23.4% have an effect size <1%. However, instead of applying thresholds to these low-effect size sQTLs, they justify their inclusion by positing that these events might be important in other developmental stages, or could produce highly toxic aberrant peptides. The figures presented in this chapter (*Figure 4.18* and *Figure 4.19*) suggest that low-effect size sQTLs are more likely to simply be noise in a dataset, as opposed to conserved events with highly deleterious effects in specific contexts.

4.4.7 *trans*-sQTLs

A common search window around a gene for *cis*-eQTL studies is 1Mbp[200]. With the advent of higher memory computing facilities, *trans*-eQTL studies have now become wide-spread[397,398], with some attributing greater influence to all cumulative *trans* effects genome-wide than individual per-gene *cis* effects. Pritchard *et al.* have posited an “omnigenic inheritance” model, with core genes directly affecting a phenotype being modulated by many *trans* effects from across the genome, and they estimate that up to 70% of heritability of a trait may be explained by *trans* as opposed to *cis* effects[399].

trans-sQTLs were attempted to be calculated for this project, however the computational requirements proved too great. It is feasible for gene-level eQTLs, however when expression of all the individual transcripts of all genes need to be analysed, the combination of each of those features with genome wide SNP variants meant that the analysis failed, even with maximum possible resources allocated from the University of Edinburgh compute cluster. The sQTLseekeR authors performed a rudimentary *trans*-sQTL analysis whereby they attempted to associate SNPs from random genes to changes in alternative splicing of distant, unrelated genes. This was intended as a test to prove that the sQTLs they identified in *cis* were more reliable and were the most likely causative SNPs. However the test actually resulted in an elementary insight into potential *trans*-sQTLs: they only found 107 significant associations between SNPs and the splicing of distant genes, however of these approximately a quarter (25) of the variants were within genes with known functions in RNA processing and transcription - implying the *trans*-sQTL

variants associated with them could truly play a role in influencing alternative splicing of distant genes[247]. *trans*-sQTLs have been identified genome-wide in *Drosophila melanogaster* by Qu *et al.*, however they have a smaller genome with fewer genes and fewer variants to assess correlations between[395]. A hotspot of *trans*-sQTLs was identified on Chr 3L, mimicking previous discoveries of *trans*-eQTL hotspots[400].

Other groups have examined somatic *trans*-sQTLs specifically in the context of cancer. Kahles *et al.* specifically considered only somatic variants arising in TCGA samples and tested them for evidence of *trans*-sQTL effects, thus greatly reducing the numbers of tests required[297]. The *trans* effects they did identify were primarily mutations in core spliceosome machinery which had wide-ranging *trans* effects on the splicing of many genes, highlighting the influence that alternative splicing can play in tumour progression.

Chapter 5 Genomic Characterisation of sQTLs

5.1 Introduction

The previous results chapter presented the distribution of the identified sQTL events in their local and genome-wide context, and set thresholds to prioritise sQTLs from more highly expressed features and with larger effect sizes. This chapter will focus on the functional relevance of the sQTLs in terms of variant effect prediction and enrichment within epigenetic marks and regions of DNA accessibility[200,247,258,260,261,264,265]. The relationship between sQTLs and eQTLs will also be analysed to determine whether sQTL variants originate independently from eQTLs; this will help to clarify whether using sQTLs in addition to eQTLs in the identification and prioritisation of CRC predisposition signals would add power[265,401]. A meta-GWAS of 58,640 individuals is used to assess whether sQTL SNPs are enriched for greater associations with CRC predisposition than would be expected by chance. Finally, a number of examples of sQTLs in genes linked to CRC predisposition or progression are given.

5.1.1 Linkage disequilibrium

As demonstrated in the previous results chapter, it is rare for a single variant to be associated with an sQTL event: most commonly multiple SNPs correlate significantly with each change in transcript expression. This is because SNPs on the same chromosome can be re-assorted via chiasmata during meiosis, with the likelihood of SNPs being separated increasing as a function of the physical distance between SNPs. This means contiguous stretches of DNA are more commonly inherited together, and so individuals from the same population may commonly inherit the same small groups of variants, making it difficult to dissect effects pertaining from different alleles of nearby SNPs because they are rarely separated in different individuals. SNPs which are more commonly co-inherited together than would be expected by chance are said to be in linkage disequilibrium (LD), and haplotype blocks are defined as groups of SNPs in high LD with each other[213,402].

A limitation of QTL studies performed using SNP array genotyping is that only a subset of variants have been surveyed, and therefore there is a non-exhaustive list of possible SNPs with which to draw associations. This complicates analysis of quantitative traits because the “lead” SNP associated most significantly with an event might not be the causative SNP, but could simply be in high linkage

disequilibrium with that variant. It is therefore important to take into account the LD structure of the sQTL variants identified in this study. When analysing the overlap of genomic features with sQTLs, the whole LD block containing the sQTL must be considered in order to have the highest likelihood of including the causative SNP.

5.1.2 Functional annotation and epigenetic states

True sQTLs should fall within functionally relevant and transcriptionally active regions of the genome[258], therefore exploring the predicted functional impacts of candidate sQTLs and the genomic environments in which they reside is important for assessing the reliability of their discovery.

Some causative variants identified as sQTLs would be predicted to have splice-relevant effects by disrupting intron-exon boundary motifs[264]. However, synonymous variants have also been observed to affect splicing and are theorised to mediate their effects by disrupting hexameric sequences located in the flanks of exons within 70bp of splice sites which enhance or suppress splicing[157,403]. These exonic splice enhancers (ESEs) and exonic splice suppressors (ESSs) act either by recruiting Serine/Arginine-rich (SR) proteins which help to recruit the U1 spliceosome complex, or by attracting heterogeneous nuclear ribonucleoproteins (hnRNPs) which inhibit binding of the U2AF spliceosome subunit to intronic branch point sequences[156].

Modifications to the tails of histone proteins can impact the compaction of chromatin[404,405] or promote the binding of proteins which influence transcription[183], and therefore have a significant effect on the expression of nearby genes[406]. The presence of such marks can be located via ChIP-seq (Chromatin ImmunoPrecipitation followed by sequencing[407]) and the accessibility or relative compaction of chromatin can be profiled by DNase sensitivity assays[408]. Germline variants do not act independently of such epigenetic marks, and their effects can be modulated by the local chromatin landscape[406]. Therefore it is important to characterise the epigenetic features in which sQTLs are located.

5.2 Data

5.2.1 Linkage disequilibrium blocks

Linkage disequilibrium blocks computed from 1000 Genomes Phase 1 Release 3 were downloaded from the EURAC consortium in March 2019:

http://www.eurac.edu/en/research/health/biomed/services/Documents/1000G_CEPH_ldblocks_wp_16102013.tar.gz

These blocks were calculated from 11 million SNPs using individuals from the CEU population (Utah individuals with European Ancestry)[409,410]. The blocks were calculated using an implementation of Gabriel's widely adopted 2002 definition of LD[213] developed by the EURAC consortium which prunes the search space to only the SNPs most likely to contribute to haplotype blocks[411]. The blocks were called by the consortium based on GRCh37 coordinates, so they were lifted-over to GRCh38 coordinates using the UCSC Liftover Tool[367] with chain file "hg19ToHg38.over.chain.gz". Only 442 of 412,206 regions were unable to be lifted over successfully: 125 were deleted in the newer assembly, 147 partially deleted and 170 split.

There were no LD blocks available from EURAC for the X chromosome, so an additional set of LD blocks was calculated using 6.6M variants for the 221 individuals from both batches using Plink version 1.9[412,413], which implements the same LD-approximation algorithm developed by the EURAC consortium[410]. The maximum block size was set to 500kbp, 2-SNP blocks were limited to a maximum of 20kbp and 3-SNP blocks to 30kbp.

There were 245,832 LD blocks inferred from the 221 Scottish patients compared to 411,605 from 1000 Genomes. The LD blocks calculated from 1000 Genomes data were consistently smaller and so of higher resolution (Table 5.1). The same overall trend of block size was comparable between the two datasets, justifying the use of the autosomal 1000 Genomes LD blocks supplemented by the X chromosome blocks generated from in-house data. Blocks for the X chromosome are likely to be larger due to the sparser coverage of SNP-genotyping.

Chr	No. blocks 1KG	No. blocks CCGG	Median kbp 1KG	Median kbp CCGG
1	30285	17137	2.29	3.16
2	33598	19552	2.16	2.82
3	27421	15936	2.16	2.66
4	26026	14656	2.06	2.96
5	24450	14380	2.23	2.83
6	25863	14357	1.67	2.48
7	24541	14161	1.83	2.53
8	21905	13260	1.78	2.34
9	19805	12267	1.70	2.05
10	20430	12092	1.91	2.46
11	20079	11612	1.92	2.62
12	19933	11330	2.03	2.75
13	14192	8348	2.10	2.81
14	13757	8013	1.88	2.56
15	13583	8033	1.75	2.29
16	15498	9593	1.37	1.73
17	12825	7365	1.72	2.32
18	11844	7542	1.98	2.36
19	12201	6981	1.30	1.84
20	9969	6437	1.86	2.21
21	6142	3789	1.59	2.33
22	7258	4186	1.42	1.92
23	NA	4805	NA	6.70

Table 5.1 LD block sizes and numbers of blocks per chromosome derived from 1000 Genomes Phase 3 release v5 (1KG) or the CCGG cohorts

5.2.2 Minor Allele Frequencies

The “Scotland Combined” cohort is an ever-growing dataset of genotypes collected by the CCG Group for the purposes of understanding CRC risk factors in the Scottish population. The size of the cohort also makes it amenable to assessing the allele frequencies of Scottish cohorts. Allele frequencies were extracted for SNPs from a data-freeze of this cohort when it contained 16,000 individuals from the Generation Scotland study[414], 1,800 from the “Scotland Phase 1” cohort[109], and 287 patients who had undergone surgery or observation as part of the SOCCS or SCOVIDS cohorts. Where the analyses in the chapter necessitated background sets of SNPs with matching allele frequencies, the sQTL SNPs were binned into 50 intervals of minor allele frequency (MAF) from 0.0 to 0.5, and 100,000 random SNPs which were non-significant for sQTLs were chosen with the same proportion from each bin, resulting in a margin of error for MAF-matching of 0.01.

When comparing the functional attributes of significant sQTL SNPs to a background set of non-significant SNPs, some groups performed LD-pruning of the background SNPs first[264]. Whilst there is merit in this approach so as to ensure that all SNPs in the background set were independent, background sets were not LD-pruned in this study because the process leaves only a single SNP per LD block. The pruning process could potentially bias the distribution of remaining SNP effects, and obscure or promote certain categorisations. Not having pruned means that there could be multiple SNPs from the same LD block contributing to the background set; however, by choosing a relatively large set of 100,000 background SNPs, the contribution of any instances of LD blocks with multiple sampled SNPs should be evenly distributed across the whole genome.

The distribution of allele frequencies from the in-house Scotland Combined cohort was compared with an external database to ensure there were no idiosyncratic biases present. Allele frequencies from the 1000 Genomes phase 3 version 5a release, using data from dbSNP release 149, were downloaded from the EBI (European Bioinformatics Institute) ftp site for the autosomes and chromosome X in March 2019: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/

The original genotyping was relative to GRCh37 coordinates, but was lifted over to GRCh38 by the EBI[409]. Allele frequencies for the EUR population specifically were extracted and compared to 6,551,177 variants from the Scotland Combined cohort, yielding a Spearman's rho correlation of 0.99 (Figure 5.1).

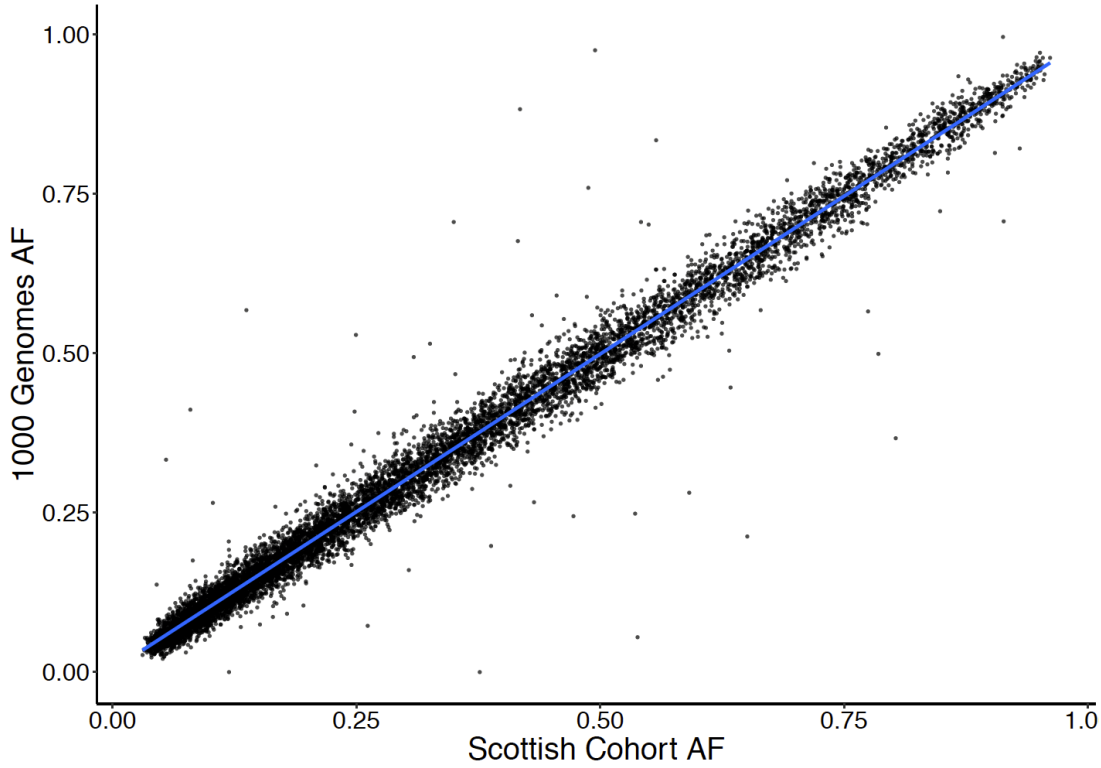


Figure 5.1 Correlation between allele frequencies from the "Scotland Combined" cohort and 1000 Genomes phase 3 release 5a EUR cohort.
A random sample of 10,000 points are plotted for clarity.

5.2.3 Functional elements and chromatin marks

The ENCODE project aims to characterise all functional elements within the human genome[415,416]. They initially studied histone modifications and DNase accessibility in human cell lines[417], and the Roadmap Epigenetics Consortium extended this to human tissue samples[418] (Table 5.2). The 2015 GTEx pilot study included an analysis of the co-localisation of quantitative trait loci within epigenetic features. They called sQTLs from nine different tissues (adipose, whole blood, blood vessel, heart, lung, muscle, nerve, skin and thyroid), and found significant enrichment of these sQTLs, and eQTLs from 43 tissues, within regulatory features identified by ChIP-seq[200].

Feature	Functional annotations	Reference
H3K4me1	Associated with enhancers and active TSSs	Barski 2007[419]
H3K4me3	Usually a strong mark of active transcription. When found within a larger region of H3K27me3, defines Bivalent/Poised promoters	Voigt 2013[420]
H3K9ac	Activating	Karmodiya 2012[421]
H3K9me3	Repressive	Lehnertz 2003[422]
H3K27ac	Activating	Tie 2009[423]
H3K27me3	Repressive	Ferrari 2014[424]
H3K36me3	A mark of actively transcribed genes, often found to demark exon boundaries	Schwartz 2009[184]

Table 5.2 ChIP-seq peaks available for colonic mucosa from the Roadmap Epigenetics Consortium[425] and their most common influences on gene expression.

In addition to individual chromatin peaks, the chromatin "state" of regions can be inferred. The Roadmap Consortium used a Hidden Markov Model, ChromHMM[426], to identify 15 different states from a combination of the 5 highest quality ChIP-seq data, annotating among other things active promoters, enhancers, repressed polycomb and heterochromatic regions[425] (Table 5.3).

Feature	Description	Average % Genome
TssA	Active TSS	0.7
TssAFlnk	Flanking active TSS	0.5
TxFlnk	Transcription at gene 5' and 3'	0.1
Tx	Strong transcription	3.6
TxWk	Weak transcription	11.6
EnhG	Genic enhancers	0.4
Enh	Enhancers	2.8
ZNF_Rpts	ZNF genes and Repeats	0.2
Het	Heterochromatin	2.6
TssBiv	Bivalent/Poised TSS	0.1
BivFlnk	Flanking Bivalent TSS/Enhancer	0.1
EnhBiv	Bivalent Enhancer	0.1
ReprPC	Repressed PolyComb	1.2
ReprPCWk	Weak Repressed PolyComb	8.3
Quies	Quiescent/Low signal region	67.8

Table 5.3 Definitions of 15 inferred Chromatin States
Percentages are the average genome coverage from 111 reference epigenomes characterised by the Roadmap Epigenetics Consortium[425].

5.2.4 Epigenetic and functional element annotations

Sample "E075" from the Roadmap Epigenetics Consortium represents colonic mucosa sampled from a 73 year old female[425]. Narrow peak bed files of ChIP-seq for the histone modifications H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3 and H3K36me3 were downloaded from the ENCODE portal[427] in March 2019: https://www.encodeproject.org/reference-epigenomes/ENCSR055HAB/ (Table 5.4, Table 5.5).

The ENCODE Consortium also predicted candidate regulatory elements based on an integrative analysis of DNase I hypersensitivity, H3K4me3, H3K27ac, and CTCF peaks[417] (Table 5.4, Table 5.5), which were downloaded for sample E075 from the ENCODE portal in March 2019: https://www.encodeproject.org/annotations/ENCSR212XWK/.

Consensus DNase I hypersensitivity data for 125 cell types was used because the DNA accessibility data for just sample E075 was of low quality (Table 5.4, Table 5.5). The ENCODE Analysis Working Group uniformly processed DNase I hypersensitivity data from teams at Duke and Washington Universities and extracted narrow peaks at a threshold of 1% FDR[428]. Clusters of DNase peaks supported by only one cell type were removed. The data was downloaded from the UCSC table browser in March 2019: http://genome-preview.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegDnaseClusteredV2.

Sample	Data content	File ID/Accession	Originally mapped to assembly
E075	H3K4me1	ENCFF908AYT	GRCh38
E075	H3K4me3	ENCFF637OKI	GRCh38
E075	H3K9ac	ENCFF647QYV	GRCh38
E075	H3K9me3	ENCFF591QMA	GRCh38
E075	H3K27ac	ENCFF051OAR	GRCh38
E075	H3K27me3	ENCFF839CHE	GRCh38
E075	H3K36me3	ENCFF745LLC	GRCh38
E075	Candidate Regulatory Elements	ENCFF952CXM	hg19
Consensus across 127 samples	DNase I hypersensitivity	wgEncodeRegDnaseClusteredV2	hg19
E075	ChromHMM	E075_15_coreMarks_dense	hg19

Table 5.4 Accession numbers of downloaded ENCODE data

Feature	n regions	Total coverage (Mbp)	Median width	Mean width
Regulatory	1309805	549.4	342	419
DNase	1281642	387.2	255	302
H3K27ac	67306	52.6	557	781
H3K36me3	107557	44.6	315	415
H3K4me3	42873	40.1	609	935
H3K4me1	74058	39.1	410	527
H3K9ac	41636	26.9	478	647
H3K27me3	48188	20.9	309	433
H3K9me3	29194	12.1	335	414

Table 5.5 Sizes of epigenetic features.
Table is ordered in descending size of total genomic region covered by each feature. Number of regions and median and mean width of the regions are presented.

The Roadmap Epigenetics Consortium used the ChromHMM tool v1.10[426] to implement a Hidden Markov model to distil 15 distinct chromatin states based on the co-occurrences of 5 different ChIP-seq peaks (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3). They used the 60 highest quality epigenomes, which included the colonic mucosa sample E075[425]. The resulting regions were annotated as characterising: active TSS (TssA), flanking active TSS (TssAFlnk), transcription at gene 5' and 3' (TxFlnk), strong transcription (Tx), weak transcription (TxWk), genic enhancers (EnhG), enhancers (Enh), ZNF genes & repeats (ZNF_Rpts), heterochromatin (Het), bivalent/poised TSS (TssBiv), flanking bivalent TSS/enhancer (BivFlnk), bivalent enhancer (EnhBiv), repressed polycomb (ReprPC), weak repressed polycomb (ReprPCWk) and quiescent/low activity regions (Quies) (Table 5.6). The corresponding regions were downloaded in March 2019 from: http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state.

Any feature files originally aligned to hg19 were were lifted-over to GRCh38 coordinates using the UCSC Liftover Tool[367] with chain file "hg19ToHg38.over.chain.gz".

Feature	n regions	Total coverage (Mbp)	Median width	Mean width
Quies	39788	2182.0	21000	54840
TxWk	55217	321.2	2800	5817
ReprPCWk	17806	143.6	4400	8064
Tx	23123	63.0	1600	2725
Enh	45922	32.7	400	712
Het	12987	25.1	1000	1933
TssA	24207	23.4	600	967
TssAFlnk	24982	12.3	400	493
ReprPC	9379	11.9	800	1264
EnhG	4878	4.1	600	837
BivFlnk	6917	3.5	400	512
ZNF_Rpts	2281	3.5	600	1535
TssBiv	5172	2.8	400	536
EnhBiv	6012	2.5	200	409
TxFlnk	1329	0.9	400	644

Table 5.6 Sizes of ChromHMM predicted chromatin states

5.2.5 GWAS associated and COSMIC genes

The NHGRI-EBI GWAS catalog[429] was queried for the trait “colorectal cancer” in September 2019 from the URL: https://www.ebi.ac.uk/gwas/efotraits/EFO_0005842

Any studies in African or Asian populations were excluded to leave only associations derived from populations with European ancestry, to best reflect the Scottish cohort used in this project.

The COSMIC catalog[430] release v90 was queried for all curated genes linked to cancer progression, including those linked to colorectal cancer specifically, in September 2019 from the URL: https://cancer.sanger.ac.uk/census

5.3 Methods

5.3.1 Variant effect prediction

The effects of sQTL SNPs were predicted using SnpEff version 4.3[431] with the SnpEff GRCh38v86 database (the most closely matched available database to the GRCh38v88 gene build used for sQTL predictions). Canonical and non-canonical transcripts were included in the predictions, and the default interval size of 5kbp was used to classify a variant as up or downstream of a given transcript. The MISO Sequence Ontology Browser[432,433] was consulted when grouping different variant effects together. The umbrella term “Splicing variant” was applied to any terms including “splice_acceptor_variant”, “splice_donor_variant”, “splice_region_variant” or “exon_loss_variant”. “splice_region_variant” is defined by the Sequence Ontology as “A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron”[432].

5.3.2 Circular permutation

Circular permutation is a method for testing enrichment of genomic features within sets of annotated regions. It involves artificially circularizing chromosomes *in silico*, and performing permutations of a test set of regions against a query set which is shifted by a random number of bases with each iteration. Circularizing the chromosomes means that biologically relevant clustering/patterns of features and the relative distances between features are maintained whilst being shifted. Circular permutations were performed with the regioneR package[434], using the UCSC hg38 genome masked for assembly gaps and intra-contig ambiguities (BSgenome.Hsapiens.UCSC.hg38.masked v1.3.99)[435].

There may be instances whereby one set of features significantly overlaps another, though only due to the second feature completely encapsulating the first. This could be caused by one set of features being significantly larger than the other, meaning that there is a large window whereby one can be shifted and still significantly overlap the other. In order to control for instances such as this, the regioneR package performs local Z-score analyses, whereby the test features are shifted a short distance either side of their native locus and the overlap with the test regions are re-analysed. If the enrichment is genuine and not simply due to inequalities between the sizes of the features, then there should be a peak of greatest Z-score

enrichments centred around the original positions of the features, which decay as the shifted distances increase.

5.3.3 Obtaining and filtering eQTLs

eQTLs were called using the 221 individuals from batches 2013152 and 10525 by Dr Victoria Svinti. Protein-coding and lncRNA genes were retained for analysis if they had at least 6 counts in at least 10% of the samples, as quantified by Salmon against GRCh38v88. Expression was TMM normalised, and rank transformation was applied to impose a normal distribution. FastQTL was run using 10 PEER factors as covariates, in addition to age, gender, and the site within the colon from which the samples were obtained. Resulting nominal p-values were FDR corrected following the methodology of the GTEx Consortium[200]. 734,788 eQTL variants passed FDR 0.05 correction, relating to 11,688 different protein-coding or lncRNA genes. Choosing the "lead" variant per gene based on the greatest significance and largest effect size left 11,039 unique variants.

eQTLs identified by the GTEx Consortium were also downloaded for transverse and sigmoid colon tissues in March 2019 from: https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz.

A considerable caveat of using eQTLs from the GTEx Consortium is that their tissue samples were obtained post mortem and the samples may have been subjected to lengthy periods of ischaemia (mean of approximately 6 hours, to as long as 21 hours)[200]. The sigmoid and transverse colon samples were found to have some of the fastest rates of decay of RNA integrity post mortem of any of the 43 tissues surveyed by GTEx[393]. The sampling method of GTEx also meant that colonic mucosa samples were prone to also including a stromal component, whereas great care was taken by the investigators obtaining samples from the Scottish cohort analysed in this study to strip the colonic mucosa from the surrounding muscle layer.

5.3.4 GWAS enrichment via lambda inflation

sQTL variants were tested for genomic enrichment within a meta-GWAS for CRC carried out by Dr Maria Timofeeva, which combined summary statistics from 10 different cohorts with European ancestry (Table 5.7). All cohorts, or subsequent

iterations thereof, have since been used as part of a larger meta-analysis by Law *et al.*[96].

Study	Curating Institute	Cases	Controls
Generation Scotland	Edinburgh University	4,551	8,804
Scotland Phase 1 Cohort	Edinburgh University	932	943
UK Biobank	Edinburgh University	1,100	3,637
VICTOR trial with 1958 birth cohort controls	Oxford University	1,794	2,686
Colorectal Tumour Gene Identification Consortium	Oxford University	890	900
Finnish Colorectal Cancer Predisposition Study	Helsinki University	1,172	8,266
National Study of Colorectal Cancer Genetics	ICR London	6,459	7,191
COIN trial	Cardiff University	1,950	2,162
Colon Cancer Family Registry Cohort 1	University of Southern California	1,175	999
Colon Cancer Family Registry Cohort 2	University of Southern California	795	2,234
	Totals	**20,818**	**37,822**

Table 5.7 Cohorts used for meta-GWAS of CRC predisposition.

Lambda inflation was calculated using the GWAS p-values corresponding to the thresholded sQTLs from each package according to Yang *et al.*'s 2011 definition, which accounts for the median expected chi-squared association of SNPs to a quantitative trait via GWAS[436].

To create null distributions, lambda inflation was calculated for the same number of SNPs as the thresholded sQTLs from each package which were able to be mapped to SNPs used in the meta-GWAS (375 for sQTLseekeR, 776 for Leafcutter) 100,000 times, randomly selecting GWAS p-values for SNPs from within the search windows of each package, and MAF-matching by increments of 0.0125. Significance was quantified via a two-tailed test on a Z-score calculated by comparing the lambda inflation score for each package's list of thresholded SNPs to the corresponding null distribution.

Expected p-values for quantile-quantile plots (QQ-plots) were calculated based on a null chi-squared distribution using the *qchisq* function from base R[437].

5.3.5 Differential Splicing Analysis

In order to generate a sashimi plot for CASP3, differential splicing was performed according the Leafcutter Authors' protocol[260], with individuals of genotype dosage 0

(n=163) compared to dosage 2 (n=5) for variant rs4647609 with no covariates added (including covariates was found to make no appreciable difference to the results). The results of the differential splicing were then visualised using the R Shiny app, LeafViz, developed by the Leafcutter Authors[260].

For visualisation in the Integrative Genomics Viewer (IGV)[333], bam files were converted to bigwig read-density tracks using the function "*bam2bw*" from the package cgpBigWig[334].

5.4 Results

5.4.1 Variant effect prediction

SnpEff variant effect predictions were made for sQTLs from sQTLseekeR and Leafcutter, and for background sets of 100,000 variants randomly selected from within the search windows of each package. Effect predictions were made for all significant sQTL SNPs and for only the sQTLs passing the expression and effect size thresholds assigned to each package in the previous chapter (Table 5.8). SnpEff consequences are assigned at the transcript-level, so multiple results were returned per variant if it fell within multiple transcripts for the same or multiple genes.

If effects are partitioned into either exonic or non-exonic (intronic, upstream, downstream, intergenic, TF binding site), there is an enrichment of sQTLs in exonic regions compared to non-exonic for significant sQTL SNPs from both packages (sQTLseekeR: two-tail Fisher's p-value= 7.74e-260, OR=2.28; Leafcutter: p-value= 4.28e-72, OR=2.34). This implies that sQTLs are more commonly located in functionally-relevant exonic regions than would be expected by chance. In both packages, the enrichment was greater for SNPs from sQTLs passing expression and effect size thresholds compared to the entire population of significant sQTL SNPs (sQTLseekeR: p-value < 2.2×10^{-16}, OR=3.62; Leafcutter: p-value= 5.94×10^{-88}, OR=2.98).

The proportion of variants falling within 5kbp up or downstream of transcripts is consistently greater for sQTLs than background SNPs in both packages, however there is a greater proportion of significant sQTL variants in intergenic regions for Leafcutter as opposed to sQTLseekeR (2.86% vs 1.34%), as would be expected given its wider search window (Table 5.8).

The enrichment or depletion of individual SnpEff annotation classes was also analysed, with p-values Bonferroni corrected for the number of different classes tested. The enrichments were of greater magnitude within thresholded sQTL SNPs than all significant sQTL SNPs (Figure 5.2). Splicing-related variants were the most highly enriched in the Leafcutter package, and among the top most enriched variants for sQTLseekeR (Figure 5.2). sQTLs were more enriched in 5' than 3' UTR regions for both packages. Although the majority of functional annotations assigned to sQTLs were intronic (44.5-57.4%, Table 5.8), this class is the least enriched relative to the background test set (Figure 5.2). Intronic SNPs were the most common and the least enriched class of variants, however lack of enrichment is not purely a function of frequency, as illustrated by intergenic SNPs which were much less common (1.34-2.86%, Table 5.8) whilst also being the second least enriched class (Figure 5.2).

a) sQTLseekeR	Significant	Significant %	Thresholded	Thresholded %	Background	Background %
Intronic	26758	52.2	1716	44.5	467724	75.9
Downstream	10849	21.1	892	23.1	64000	10.4
Upstream	10389	20.2	873	22.7	63271	10.3
Non-coding	799	1.56	91	2.36	3862	0.627
Intergenic	687	1.34	71	1.84	6777	1.1
3' UTR	644	1.26	71	1.84	3463	0.562
5' UTR	347	0.676	41	1.06	1105	0.179
Synonymous	268	0.522	36	0.934	1281	0.208
Splicing variant	176	0.343	17	0.441	703	0.114
Other	174	0.339	7	0.182	2955	0.48
Missense/Frameshift/INDEL	173	0.337	31	0.804	652	0.106
TF binding site	35	0.0682	2	0.0519	220	0.0357
Stop gained/retained/lost	7	0.0136	6	0.156	10	0.00162
b) Leafcutter	**Significant**	**Significant %**	**Thresholded**	**Thresholded %**	**Background**	**Background %**
Intronic	40268	57.4	3866	56.2	288041	67.2
Upstream	12013	17.1	1116	16.2	47834	11.2
Downstream	11885	16.9	1250	18.2	47779	11.1
Intergenic	2007	2.86	155	2.25	34087	7.95
Non-coding	1204	1.71	145	2.11	3458	0.807
3' UTR	728	1.04	50	0.727	2586	0.603
5' UTR	586	0.835	67	0.974	938	0.219
Splicing variant	464	0.661	102	1.48	504	0.118
Synonymous	398	0.567	72	1.05	932	0.217
Other	360	0.513	21	0.305	1813	0.423
Missense/Frameshift/INDEL	243	0.346	33	0.48	499	0.116
TF binding site	48	0.0684	2	0.0291	234	0.0546
Stop gained/retained/lost	3	0.00427	0	0.0	11	0.00257

Table 5.8 Number and percentage of variant effects assigned by SnpEff to all significant sQTLs, thresholded significant sQTLs or 100,000 randomly selected background lists of sQTLs from the search windows of sQTLseekeR or Leafcutter

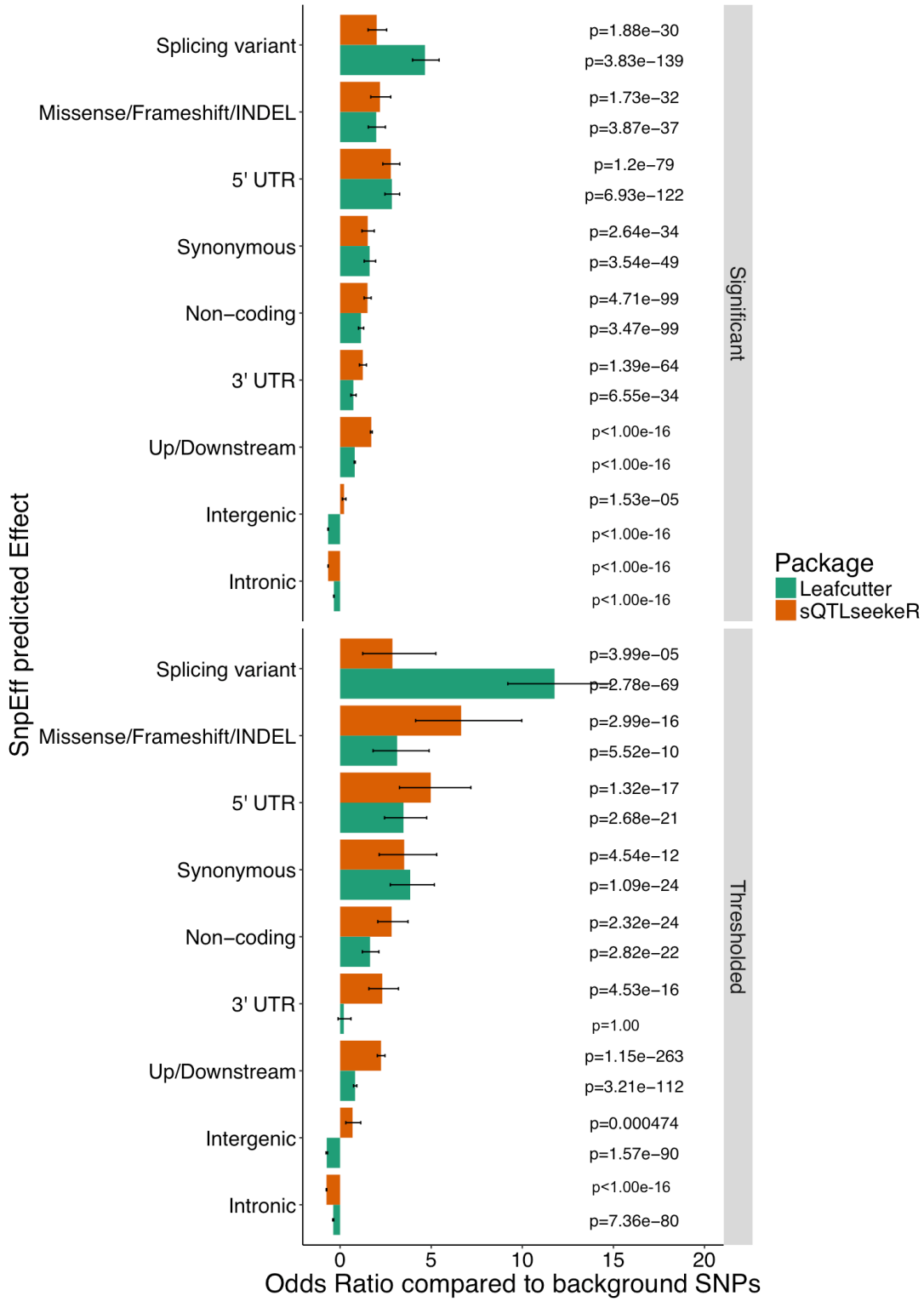


Figure 5.2 Enrichment of SnpEff consequence classes
Enrichment calculated between either all significant sQTLs or thresholded sQTLs and a set of 100,000 randomly chosen, MAF-matched SNPs from within the search window of each package. Enrichments calculated by two-tailed Fisher's test.

5.4.2 Enrichment of sQTLs in epigenetic and functional annotations

The thresholded sQTLs from both sQTLseekeR and Leafcutter were combined together and assigned to the set of LD blocks created by combining autosomes from 1000 Genomes and X chromosome from the 221 Scottish individuals. The 1,280 SNPs fell within 965 different LD blocks. The LD block coordinates were circularly permuted against 7 different ChIP-seq peaks: candidate regulatory regions predicted from Epigenetics Roadmap colonic mucosa sample E075, and consensus DNase I hypersensitivity peaks from 125 ENCODE cell types. Z-scores of enrichment were calculated by comparing the observed number of overlaps between the original positions of sQTL LD blocks and the given features against a distribution of overlaps obtained from 10,000 circular permutations of the features. Nominal p-values derived from the z-scores were Bonferroni-corrected.

Significant enrichments were found relative to ChIP-seq peaks H3K4me1, H3K4me3, H3K9ac, H3K27ac and H3K36me3, as well as the predicted candidate regulatory regions and consensus DNase I hypersensitivity sites (Table 5.9 and Figure 5.3). There was a nominally significant depletion in ChIP-seq peaks of H3K9me3 which did not survive multiple testing correction, and the number of overlaps observed against H3K27me3 peaks did not significantly deviate from the random permutations. Local Z-score plots demonstrate that the majority of the significant associations are dependent on the specific position of the features as demonstrated by a pronounced peak or valley of Z scores (Figure 5.4).

Feature	Observed overlaps	Z score	Nominal p-value	Alternative hypothesis	Bonferroni p-value
H3K4me1	337	13.20	1.00E-04	greater	9.00E-04
H3K4me3	339	17.35	1.00E-04	greater	9.00E-04
H3K9ac	304	17.33	1.00E-04	greater	9.00E-04
H3K9me3	73	-2.04	2.31E-02	less	2.08E-01
H3K27ac	366	17.14	1.00E-04	greater	9.00E-04
H3K27me3	91	-0.88	2.04E-01	less	1.00
H3K36me3	396	21.16	1.00E-04	greater	9.00E-04
Regulatory	849	8.19	1.00E-04	greater	9.00E-04
DNase	855	9.35	1.00E-04	greater	9.00E-04

Table 5.9 Significance of overlaps between combined thresholded sQTLs and individual ChIP-seq peaks, predicted regulatory regions and DNase I hypersensitivity sites. *Observed overlaps of 965 sQTL-containing LD-blocks. Z-score and p-value relative to 10,000 circular permutations of features using with respect to stated alternative hypothesis.*

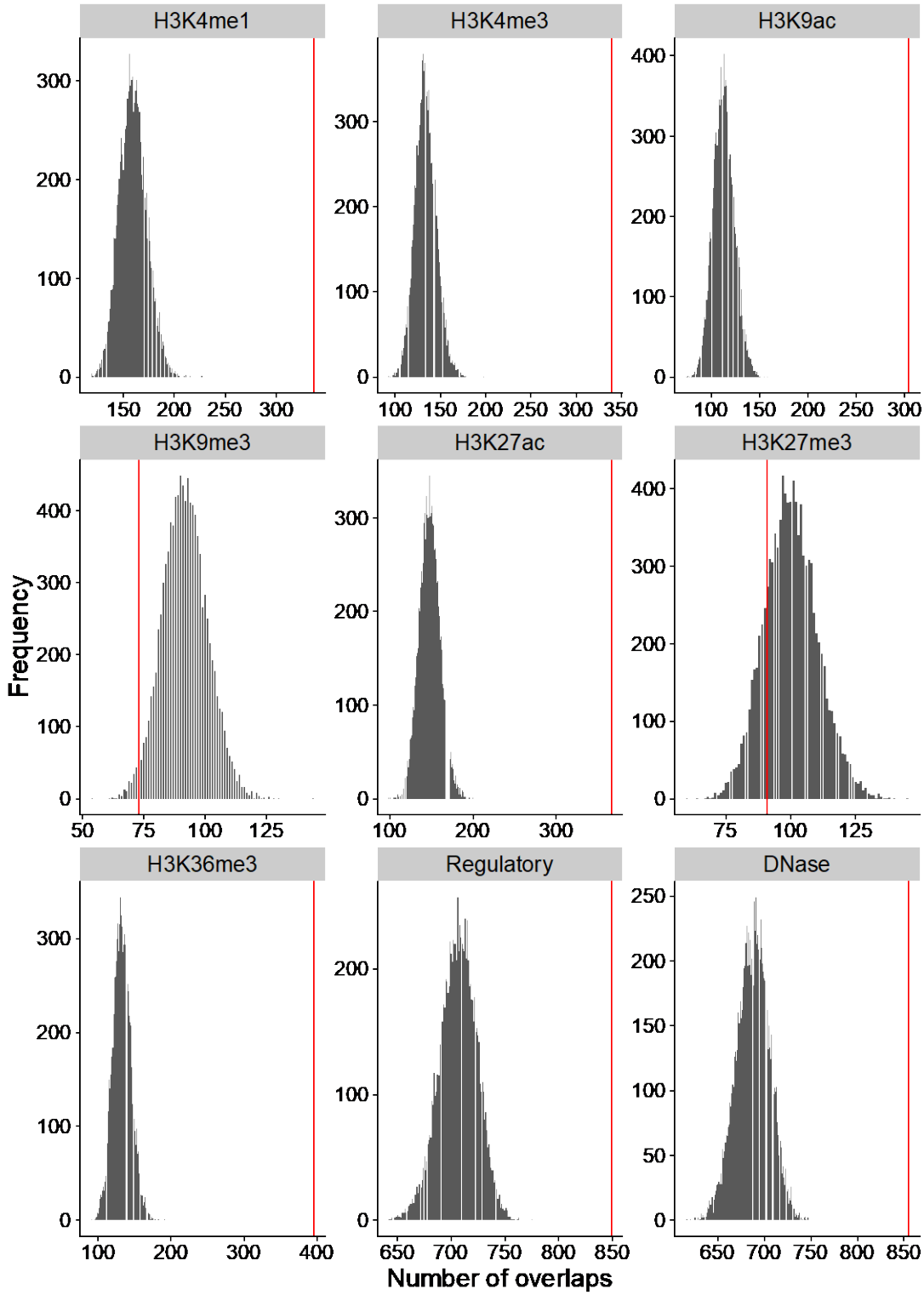


Figure 5.3 Distribution of numbers of overlaps between random permutations of ChIP-seq peaks, predicted regulatory regions and DNase I hypersensitivity sites and LD blocks containing combined thresholded sQTLs from sQTLseekeR and Leafcutter. *Grey bars represent the null distributions obtained from 10,000 random permutations of features. Red lines indicate the number of overlaps between the original feature positions and LD blocks containing sQTLs.*

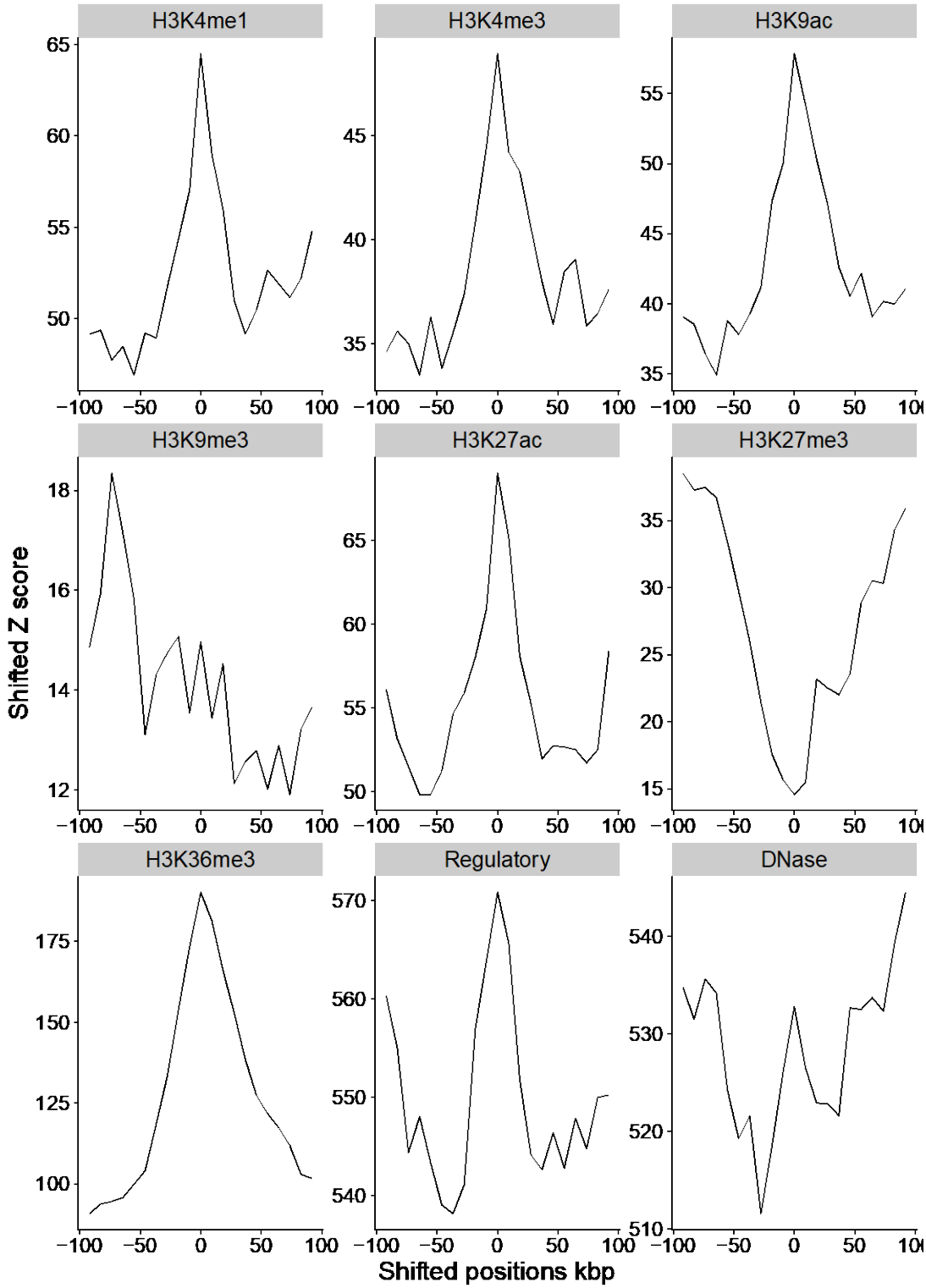


Figure 5.4 Local Z scores calculated by shifting permuted features up to +/-100kbp when overlapping ChIP-seq peaks, predicted regulatory regions and DNase I hypersensitivity sites with LD blocks containing combined thresholded sQTLs.

The combined sQTLseekeR and Leafcutter thresholded sQTLs were also tested for enrichment within the 15 chromatin states identified by ChromHMM, with Bonferroni

correction for the number of states tested. There were significant enrichments of sQTLs within all markers of active transcription and enhancer regions, and significant depletion of sQTLs within regions marked as quiescent (Table 5.10 and Figure 5.5). There was a depletion of sQTLs in heterochromatic regions which approached but did not pass significance, and there were no significant enrichments or depletions in bivalent/poised TSSs, repressed polycomb or weakly repressed polycomb (Table 5.10). The local Z scores again showed the expected peaks and troughs corresponding to enrichments or depletions (Figure 5.6).

Feature	Observed overlaps	Z score	p-value	Alternative hypothesis	Bonferroni p-value
TssA	284	19.19	1.00E-04	greater	1.50E-03
TssAFlnk	227	16.25	1.00E-04	greater	1.50E-03
TxFlnk	38	10.48	1.00E-04	greater	1.50E-03
Tx	306	21.68	1.00E-04	greater	1.50E-03
TxWk	535	19.06	1.00E-04	greater	1.50E-03
EnhG	93	14.60	1.00E-04	greater	1.50E-03
Enh	292	13.42	1.00E-04	greater	1.50E-03
ZNF_Rpts	36	6.93	1.00E-04	greater	1.50E-03
Het	47	-1.67	5.09E-02	less	7.63E-01
TssBiv	26	0.84	2.26E-01	greater	1.00
BivFlnk	35	1.75	5.42E-02	greater	8.13E-01
EnhBiv	32	2.21	2.48E-02	greater	3.72E-01
ReprPC	33	-0.04	5.33E-01	less	1.00
ReprPCWk	101	-0.18	4.69E-01	less	1.00
Quies	532	-10.10	1.00E-04	less	1.50E-03

Table 5.10 Significance of overlaps between combined thresholded sQTLs and 15 chromatin states predicted by ChromHMM.

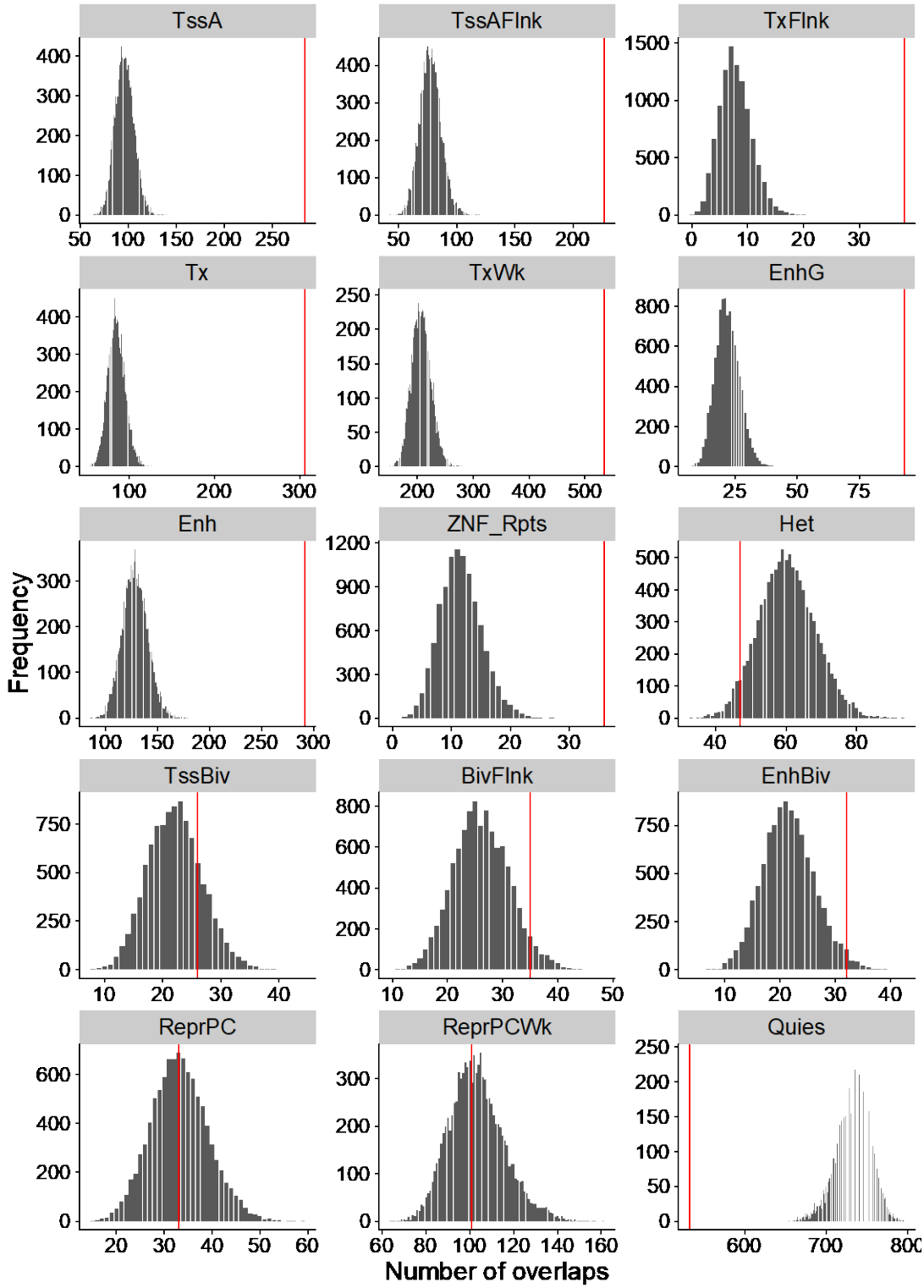


Figure 5.5 Distribution of numbers of overlaps between random permutations of 15 predicted chromatin states and LD blocks containing combined thresholded sQTLs from sQTLseekeR and Leafcutter.
Grey bars represent the null distributions obtained from 10,000 random permutations of chromatin states. Red lines indicate the number of overlaps between the original states and LD blocks containing sQTLs.

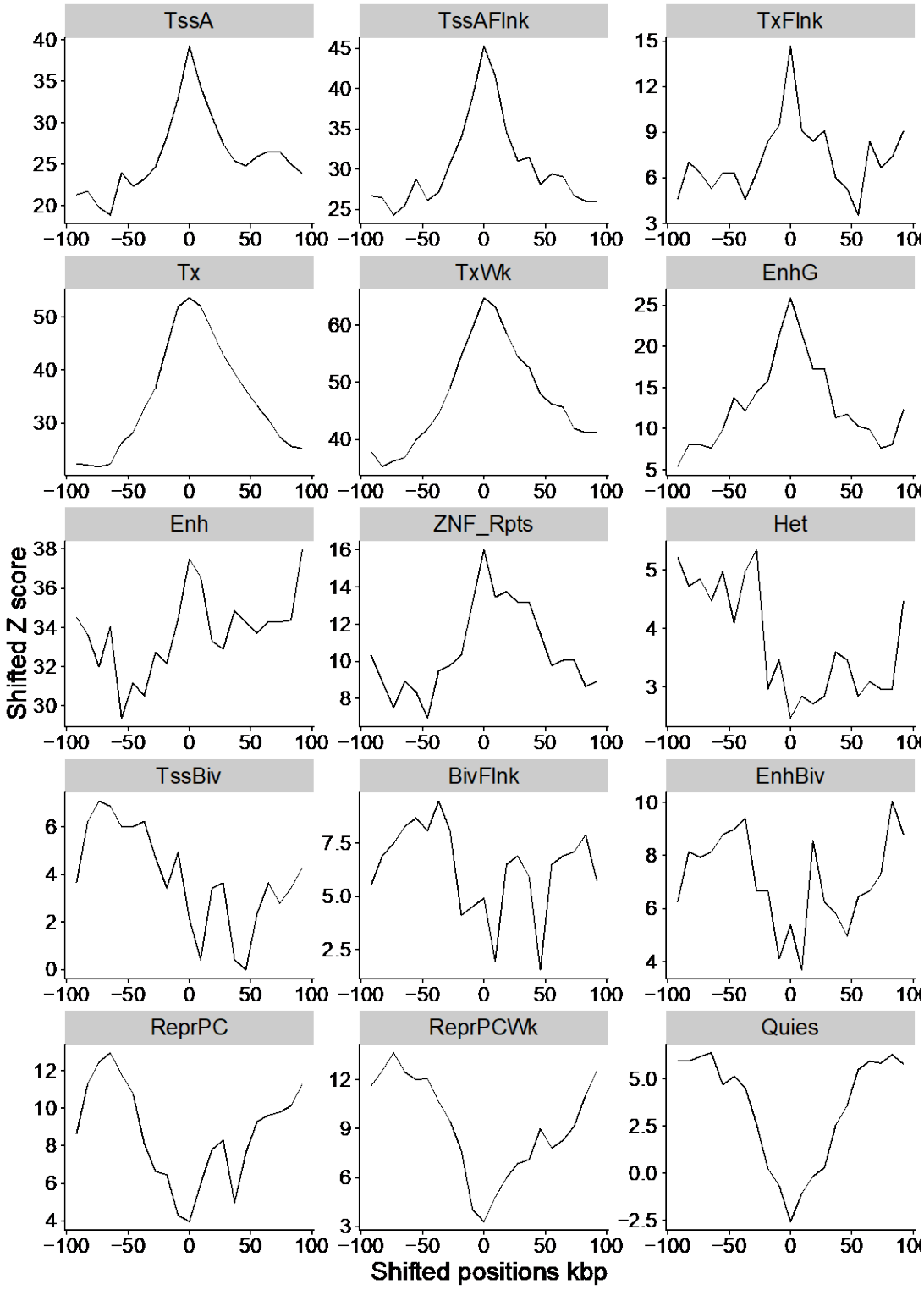


Figure 5.6 Local Z scores calculated by shifting permuted states up to +/-100kbp when overlapping 15 chromatin states with LD blocks containing combined thresholded sQTLs.

5.4.3 Relationship between sQTLs and eQTLs

68.8-71.3% of genes for which there were sQTLs also had eQTLs associated with them. However, few of the lead sQTLs, either per feature (per transcript-pair for sQTLseekeR or per intron for Leafcutter) or per gene, were also a lead eQTL SNP (2.14-2.79%, Table 5.11).

Comparison	sQTLseekeR	Leafcutter	Combined
Lead SNPs per Feature (%)	117 (2.14)	229 (2.26)	332 (2.16)
Lead SNPs per Gene (%)	95 (2.79)	101 (2.54)	188 (2.58)
Total SNPs (%)	49328 (51.6)	4768 (47.0)	52947 (51.0)
Genes (%)	2380 (69.6)	2873 (71.3)	4148 (68.8)

Table 5.11 sQTLs which are also eQTLs.
Numbers of sQTLs which are also eQTLs for: the lead sQTL SNP per feature (per transcript-pair for sQTLseekeR and per intron for Leafcutter); the lead sQTL SNP per gene; all FDR significant SNPs; and genes which have QTL events. sQTLs tested were from sQTLseekeR, Leafcutter or a combination of the two.

LD blocks containing the combined sQTLseekeR and Leafcutter thresholded sQTLs were circularly permutated against LD blocks containing the eQTLs derived from the Scottish dataset, and also against eQTLs called by the GTEx Consortium from sigmoid colon and transverse colon tissues[4]. When the two sets of QTLs were assigned to the same set of LD blocks derived from the 1000 Genomes Phase 1 Release 3 CEPH population, the circular permutations produced significant overlaps against all 3 selections of eQTLs (Table 1.12). However, there was an unexpected trough observed at the centre of plots tracing the Z-scores when coordinates of the LD blocks were locally shifted (Figure 1.7). This likely resulted from permuting the same set of LD blocks against one other.

Feature	Observed overlaps	Z score	p-value	Alternative hypothesis	Bonferroni p-value
eQTLs CCGG	257	9.9491	1.00E-04	greater	3.00E-04
GTEx Sigmoid	167	10.7213	1.00E-04	greater	3.00E-04
GTEx Transverse	212	14.0646	1.00E-04	greater	3.00E-04

Table 5.12 Significance of overlaps between eQTLs and sQTLs assigned to same set of LD blocks

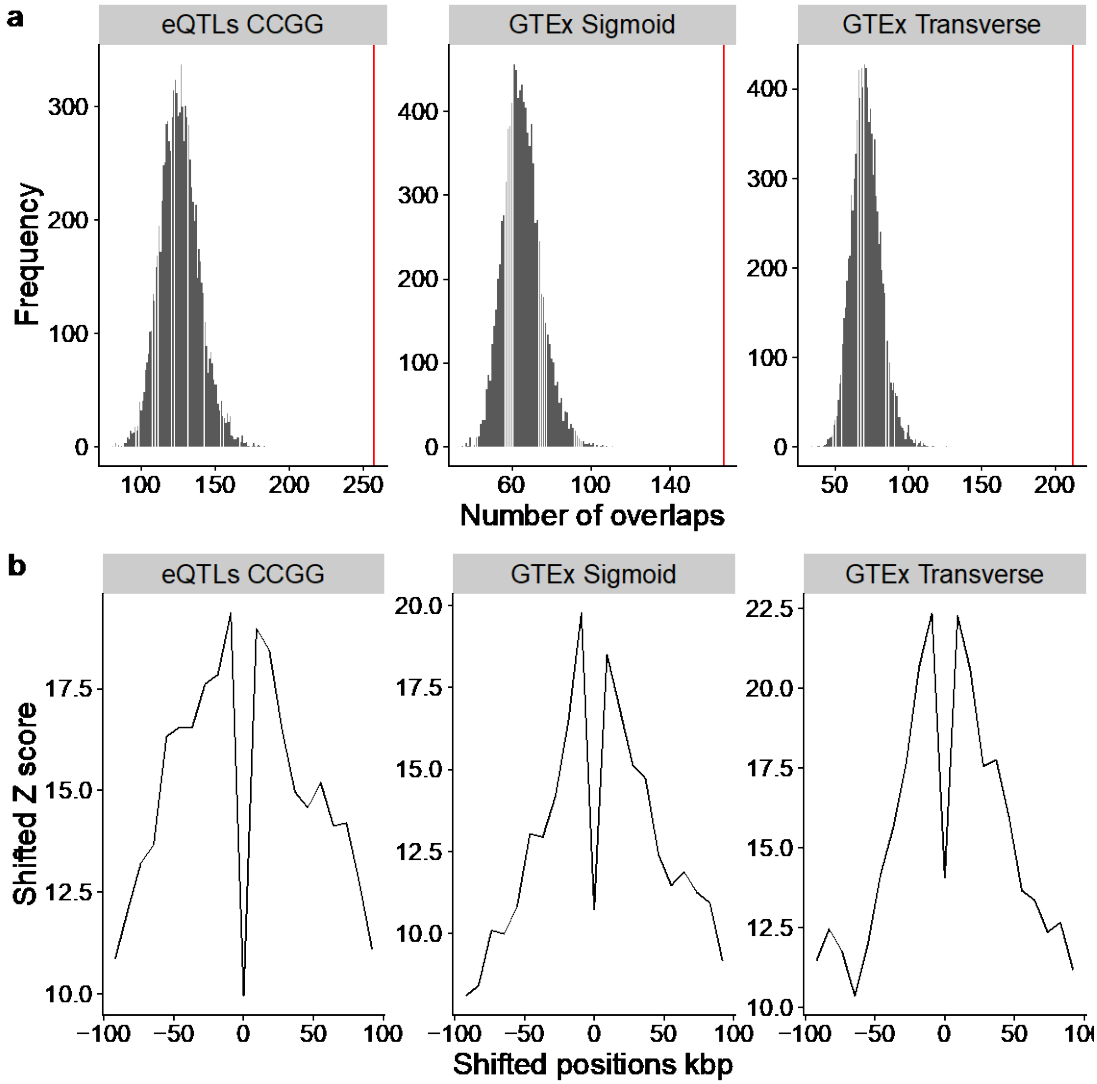


Figure 5.7 Circular Permutations of sQTLs and eQTLs assigned to same LD blocks ***a)*** *Distribution of numbers of overlaps between random permutations of LD blocks containing eQTLs from the Scottish cohort or from GTEx sigmoid or transverse colon tissues and LD blocks containing combined thresholded sQTLs. Grey bars represent the null distributions obtained from 10,000 random permutations of eQTLs, red lines indicate the number of overlaps between the original LD blocks containing eQTLs and sQTLs.* ***b)*** *Local Z scores calculated by shifting permuted eQTL LD blocks up to +/-100kbp when overlapping LD blocks containing eQTLs and sQTLs.*

Circular permutations were therefore run a second time with sQTLs assigned to the LD blocks derived directly from the genotypes of the 221 individuals in the Scottish cohort. These LD blocks were derived from a relatively small population, and so are consistently larger than the LD blocks derived from 1000 Genomes (Table 5.1). It is not ideal to use LD blocks derived from a specific population to then test hypotheses in that same population, however it was considered justified in this instance to use an available set of different LD blocks in order to re-assess the circular

permutations. The sQTLs were chosen to be assigned to LD blocks from the Scottish cohort in order to minimise the impact of using larger blocks, because there were fewer sQTLs than eQTLs being tested. As there were no LD blocks available for the X chromosome from the 1000 Genomes dataset, it was not included. There were once again significant overlaps of LD block containing sQTLs with LD blocks containing eQTLs (Table 5.13), and the local Z-scores from these permutations displayed the expected central peaks, indicating that the specific location of the features being tested produced the significance (Figure 5.8).

Feature	Observed overlaps	Z score	p-value	Alternative hypothesis	Bonferroni p-value
eQTLs CCGG	442	16.90	1.00E-04	greater	3.00E-04
GTEx Sigmoid	306	18.46	1.00E-04	greater	3.00E-04
GTEx Transverse	346	19.88	1.00E-04	greater	3.00E-04

Table 5.13 Significance of overlaps between sQTLs and eQTLs assigned to different sets of LD blocks

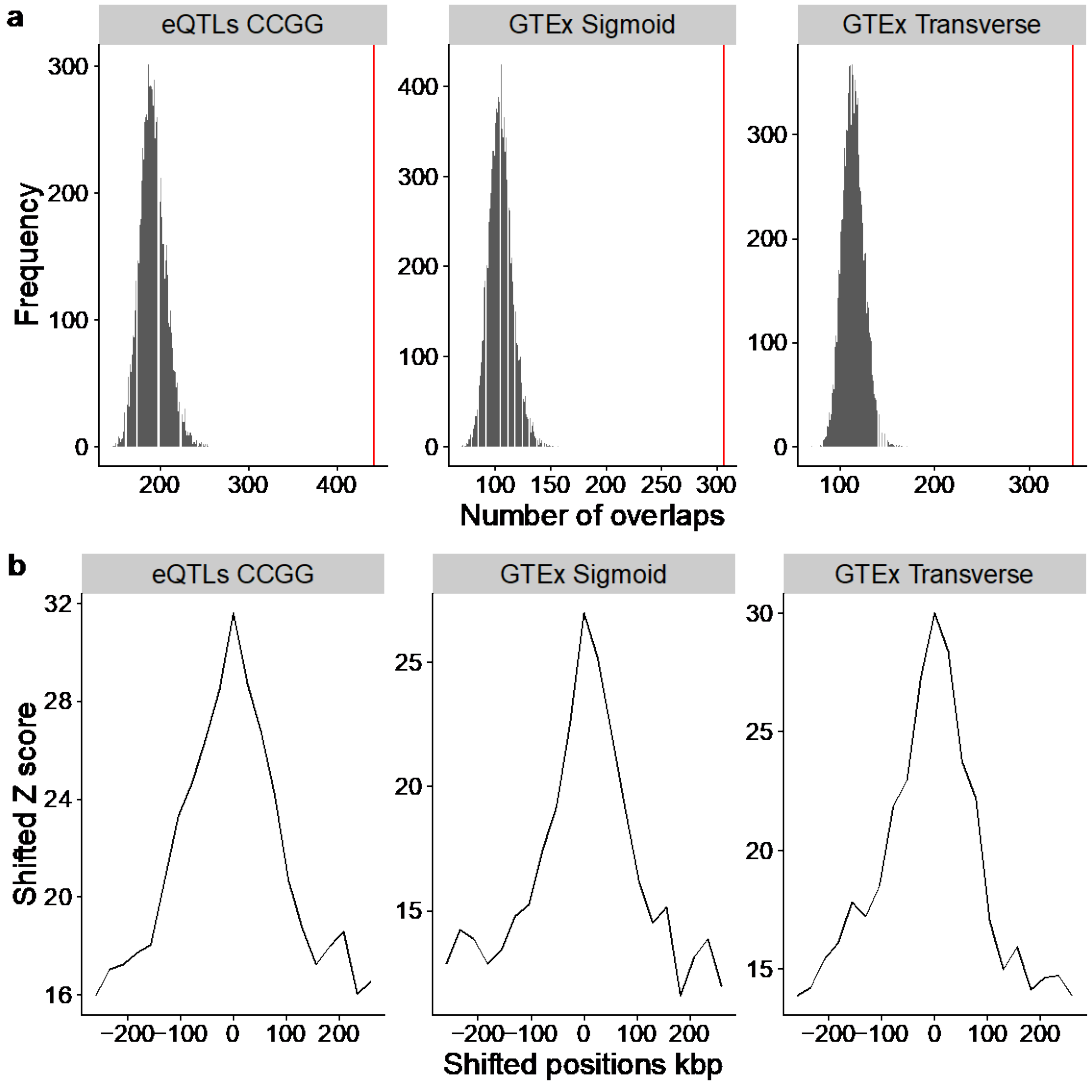


Figure 5.8 Circular Permutations of sQTLs and eQTLs assigned to different LD blocks ***a)*** *Distribution of numbers of overlaps between random permutations of LD blocks containing eQTLs from the Scottish cohort or from GTEx sigmoid or transverse colon tissues and LD blocks containing combined thresholded sQTLs. Grey bars represent the null distributions obtained from 10,000 random permutations of eQTLs, red lines indicate the number of overlaps between the original LD blocks containing eQTLs and sQTLs.* ***b)*** *Local Z scores calculated by shifting permuted eQTL LD blocks up to +/-100kbp when overlapping LD blocks containing eQTLs and sQTLs.*

5.4.4 GWAS enrichment via lambda inflation

The null distribution of lambda inflation calculated using meta-GWAS p-values from SNPs within the sQTLseekeR search window had a mean of 1.243 and median of 1.236. The lambda inflation of the 375 thresholded sQTLseekeR sQTLs using p-values from the meta-GWAS for CRC was 1.305, which produced a Z-score of

0.408 and two-tailed p-value of 0.682 (Figure 5.9). The null distribution for Leafcutter had a mean and median lambda inflation of 1.278 and 1.274, compared to the score for the 776 thresholded sQTL SNPs of 1.595, which produced a Z-score of 2.996 and a p-value of 0.00273 (Figure 5.9).

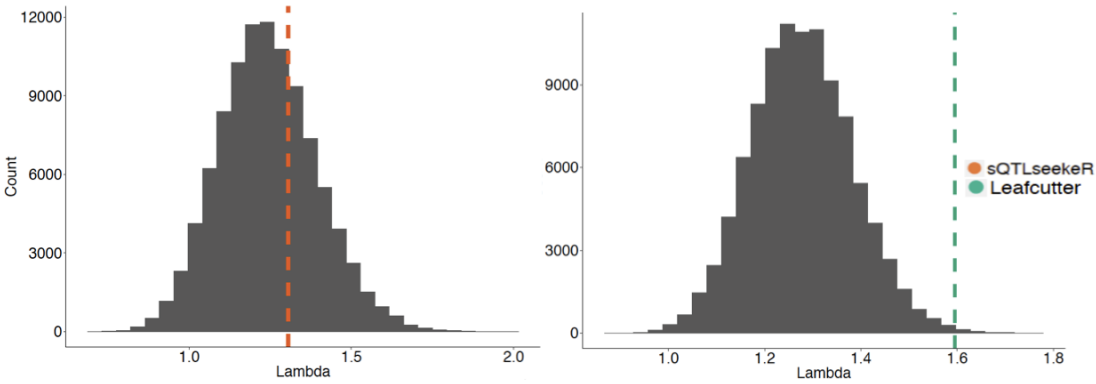


Figure 5.9 Lambda inflation distributions from 100,000 SNPs MAF-matched and selected from the same search windows as sQTLseekeR and Leafcutter

Despite not being significantly enriched compared to their corresponding background distribution of lambda inflation scores, there were sQTL variants identified by sQTLseekeR which deviated markedly from the null distribution of expected quantiles of meta-GWAS p-values (Figure 5.10).

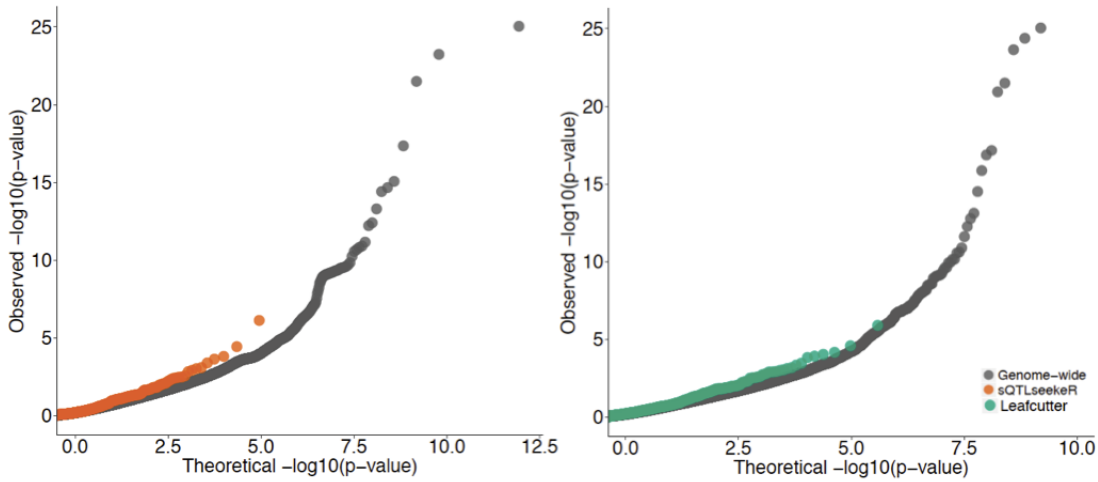


Figure 5.10 Observed against expected CRC meta-GWAS p-values for sQTL SNPs and SNPs from within the search windows of the respective packages.
For clarity of plotting, 100,000 of the genome-wide variants from the meta-GWAS were chosen, equally stratified across the quantiles of p-values.

5.4.5 sQTLs in GWAS-implicated and COSMIC genes

sQTLseekeR identified sQTLs in three genes which have been associated with CRC via GWAS (Table 5.14), and Leafcutter identified six (Table 5.15).

sQTLseekeR identified an sQTL for the *SCG5* gene which involved a 44% change in mean relative transcript expression ratio. ENST00000413748, the 211 amino acid canonical protein-coding transcript, reduced in expression relative to ENST0000030015, which has a different 5' splice site of exon 3 leading to the inclusion of an extra alanine (Figure 5.11). ENST00000498069, a 612bp non protein-coding lncRNA, remained stable at a low level in relation to the genotype of rs72715244.

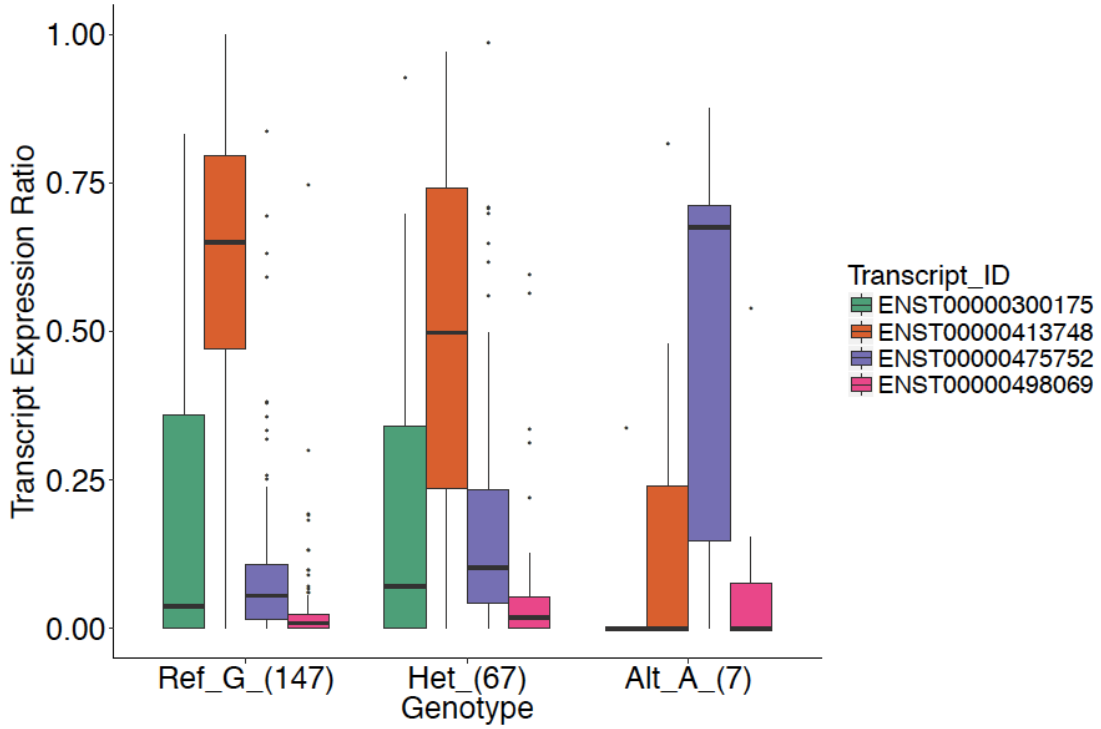


Figure 5.11 Change in SCG5 transcript expression ratio. *Reference and alternative alleles are shown, and numbers of individuals of each genotype are detailed in brackets.*

sQTLseekeR identified sQTLs in 13 genes recorded in the COSMIC catalog[430] as associated with the progression of cancer, including *PTPRT* which has been specifically linked to CRC (Table 5.16). Leafcutter identified 23 COSMIC genes, including *EIF3A*, *POLE*, and *SIRPA* which have been observed to be recurrently mutated in CRC (Table 5.17).

Gene	sQTL SNP	MAF 1KG	SnpEff variant effect	MD	FDR q-value	Mean counts	NHGRI p-value	OR (95% CI)	References
SCG5	rs72715244	0.13	intronic	0.44	2.13E-05	195.9	3.00E-27 2.00E-08	1.22 [1.18-1.27] 1.18 [1.11-1.24]	Law *et al.*[96] Peters *et al.*[438]
HLA-DQA1	rs9272330	0.57	intronic	0.51	2.13E-05	2196.3	4.00E-08	1.08 [1.05-1.11]	Law *et al.*[96]
C8orf37-AS1	rs4236835	0.12	intronic	0.21	1.23E-02	25.0	4.00E-06	1.27 [1.14-1.40]	Fernandez-Rozadilla *et al.*[439]

Table 5.14 Genes associated with CRC from the NHGRI-EBI catalog for which there were sQTLseekeR sQTLs passing expression and effect size thresholds. *"Mean counts": mean Salmon counts for each gene across all 221 primary samples.*

Gene	sQTL SNP	MAF 1KG	SnpEff variant effect	2 * slope	FDR q-value	Mean counts	NHGRI p-value	OR (95% CI)	References
HLA-DQA1	rs41268942	0.17	3' UTR	-3.24	1.85E-35	607.1	4.00E-08	1.08 [1.05-1.11]	Law *et al.*[96]
HLA-DRB1	rs2395516	0.31	intergenic	-2.46	1.83E-17	499.0	4.00E-08	1.08 [1.05-1.11]	Law *et al.*[96]
FEN1	rs61897793	0.15	intronic	-2.44	1.44E-18	93.9	1.00E-06	1.06 [1.03-1.08]	Law *et al.*[96]
PLEKHG6	rs12828469	0.62	intronic	2.52	1.77E-38	15.4	1.00E-10	1.12 [1.08-1.16]	Law *et al.*[96]
FADS2	rs7943728	0.15	intronic	2.26	1.83E-19	33.7	1.00E-06	1.06 [1.03-1.08]	Law *et al.*[96]
ERAP1	rs26500	0.79	intronic	-2.36	7.00E-12	378.5	7.00E-08	1.51 [1.23-1.86]	Al-Tassan *et al.*[227]

Table 5.15 Genes associated with CRC from the NHGRI-EBI catalog for which there were Leafcutter sQTLs passing expression and effect size thresholds. *"Mean counts": mean number of reads aligned to each intron inferred by Leafcutter across all 221 primary samples.*

Gene	sQTL SNP	MAF 1KG	SnpEff variant effect	MD	FDR q-value	Mean counts	Tumour types
PTPRT	rs6030443	0.19	intronic	0.29	3.59E-02	102.1	colorectal, HNSCC, gastric, lung, melanoma
*ERBB4**	rs10192485	0.57	intronic	0.27	6.58E-03	28.2	melanoma, gastric, NSCLC
CASP3	rs200872527	0.02	1bp intronic insertion	0.29	2.13E-05	4924.4	ovarian
COX6C	rs10542429	0.14	intronic	0.42	2.13E-05	8601.0	uterine leiomyoma
PAX8	rs3748915	0.16	intronic	0.40	2.13E-05	759.5	follicular thyroid
CNOT3	rs36634	0.44	intronic	0.26	2.13E-05	605.8	T-ALL
LPP	rs4686480	0.29	intronic	0.29	2.13E-05	51218.4	lipoma, leukaemia
MUC4	rs11922145	0.15	intronic	0.35	2.40E-02	1518.7	HNSCC
TRIM27	rs929042	0.26	intronic	0.24	2.13E-05	1050.0	papillary thyroid
PAX3	rs72960894	0.11	intronic	0.49	5.86E-05	27.1	alveolar rhabdomyosarcoma
CNTNAP2	rs10215201	0.34	intronic	0.25	2.13E-05	716.9	glioma, melanoma
MYCN	rs34039085	0.49	4bp deletion	0.26	1.47E-02	31.8	neuroblastoma, Wilms tumour
ECT2L	rs1157388	0.35	intronic	0.20	1.56E-02	254.0	ETP-ALL

Table 5.16 Genes in the COSMIC database for which there was an sQTL identified by sQTLseekeR. ** denotes the gene has also been linked to germline predisposition to cancer. ETP-ALL: early T-cell precursor acute lymphoblastic leukaemia, HNSCC: head and neck squamous cell carcinoma, NSCLC: non small cell lung cancer, T-ALL: T-cell acute lymphoblastic leukaemia.*

Gene	sQTL SNP	MAF 1KG	SnpEff variant effect	2 * slope	FDR q-value	Mean counts	Tumour types
EIF3E	rs674391	0.60	intronic	2.26	1.56E-23	38.5	colorectal
*POLE**	rs4077170	0.70	NMD	2.62	6.23E-33	73.9	colorectal, endometrioid, stomach, skin
SIRPA	rs56301259	0.37	intronic	-2.40	6.88E-33	34.9	HNSCC, colorectal, lung SCC
*FANCA**	rs12925427	0.32	intronic	-2.78	2.02E-32	34.5	AML, leukaemia
*FEN1**	rs61897793	0.15	intronic	-2.44	1.44E-18	93.9	breast
LPP	rs9877579	0.30	intronic	3.20	5.73E-77	177.9	lipoma, leukaemia
COX6C	rs34830464	0.18	intronic	-3.22	5.00E-43	39.7	uterine leiomyoma
ZNF429	rs59654184	0.20	intronic	3.52	1.18E-60	11.8	GBM
ACSL3	rs6726737	0.34	non-coding	2.80	7.80E-45	129.3	prostate
EML4	rs10490555	0.31	intronic	-2.74	3.03E-35	321.1	NSCLC
MUC1	rs2974937	0.55	intronic	2.98	3.16E-75	28.2	B-NHL
ITGAV	rs9333290	0.30	intronic	2.98	1.13E-66	6.8	large intestine carcinoma
NFE2L2	rs10930786	0.82	intronic	4.24	1.31E-70	15.5	NSCLC, HNSCC
SS18	rs994729	0.27	upstream	-2.42	1.93E-28	161.5	synovial sarcoma
CREB3L2	rs66593747	0.64	inframe deletion	-2.60	2.96E-38	62.4	fibromyxoid sarcoma
CANT1	rs12452918	0.11	intronic	-2.32	6.37E-17	27.4	prostate
CASP3	rs4647609	0.15	intronic	2.40	7.13E-17	80.7	ovarian
DROSHA	rs17409803	0.26	intronic	-2.40	8.20E-35	11.9	Wilms tumour, NSCLC, bladder carcinoma
NT5C2	rs1163248	0.77	intronic	-2.76	1.32E-29	18.7	relapse ALL
NCOR2	rs1244053	0.14	intronic	-2.28	8.70E-14	29.0	prostate
A1CF	rs12254249	0.28	intronic	-2.38	1.03E-26	9.8	melanoma
NUP98	rs12271649	0.13	intronic	2.26	2.00E-10	11.0	AML
AKAP9	rs6950470	0.39	intronic	-2.36	2.27E-28	6.3	papillary thyroid

Table 5.17 Genes in the COSMIC database for which there was an sQTL identified by Leafcutter. ** denotes the gene has also been linked to germline predisposition to cancer. ALL: acute lymphocytic leukaemia, AML: acute myeloid leukaemia, B-NHL: B-cell non-Hodgkin lymphoma, GBM: glioblastoma multiforme, lung SCC: lung squamous cell carcinoma.*

In relation to rs6030443, there was a reciprocal 29% change in transcript expression ratio within *PTPRT*, with the 31 exon transcript ENST00000373109 decreasing whilst transcript ENST00000612229, which lacks the last 7 exons, increased (Figure 5.12). Transcript ENST00000373201, which has a single 9 amino acid exon fewer than ENST00000373109, was also decreased to a lesser extent in relation to rs6030443, whilst transcript ENST00000620410, which contains only 5 exons, remained stable.

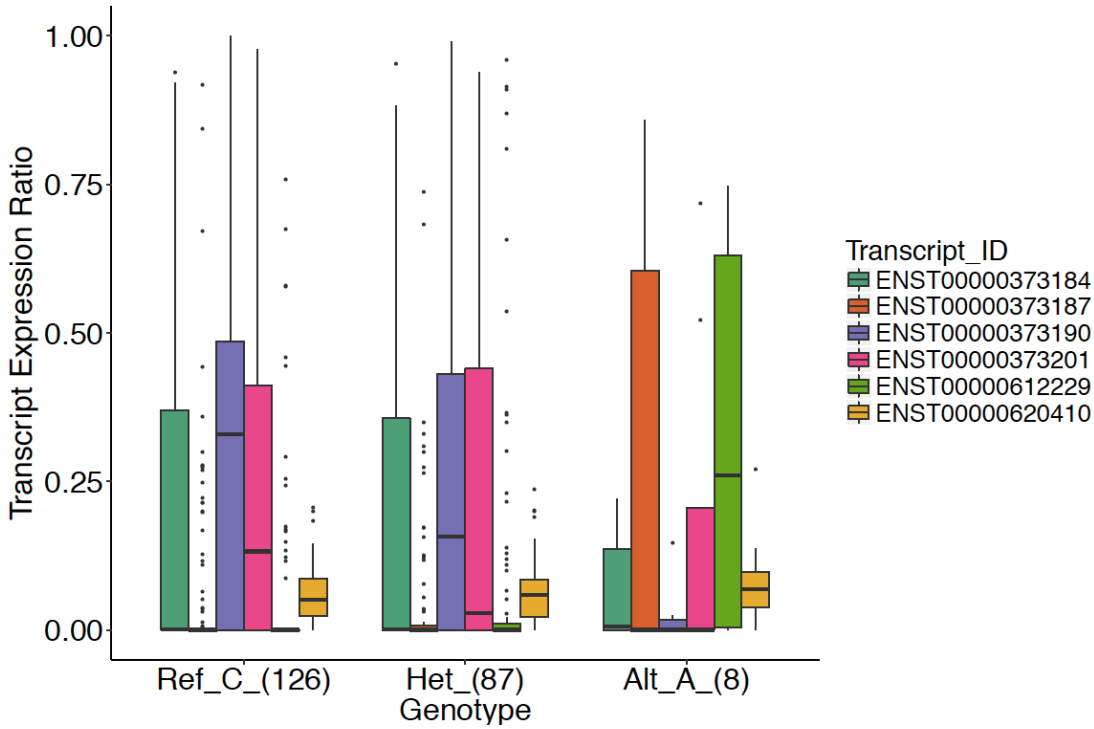


Figure 5.12 Change in PTPRT transcript expression ratio. *Reference and alternative alleles are shown, and numbers of individuals of each genotype are detailed in brackets.*

The COSMIC database contains curated associations between somatic and germline mutations in *ERBB4* and a number of cancers including melanoma, gastric and NSCLC[430]. It has not yet been recorded in the database as associated with CRC, though examples of this link are present in the wider medical literature[440,441]. In relation to the variant rs10192485, the 1,266 amino acid transcript ENST00000402597 decreased by an average of 27% relative to ENST00000436443, which includes an extra 26 amino acid final exon (Figure 5.13). The 5 exon non-coding lncRNA ENST00000484474 remained stable at a low level.

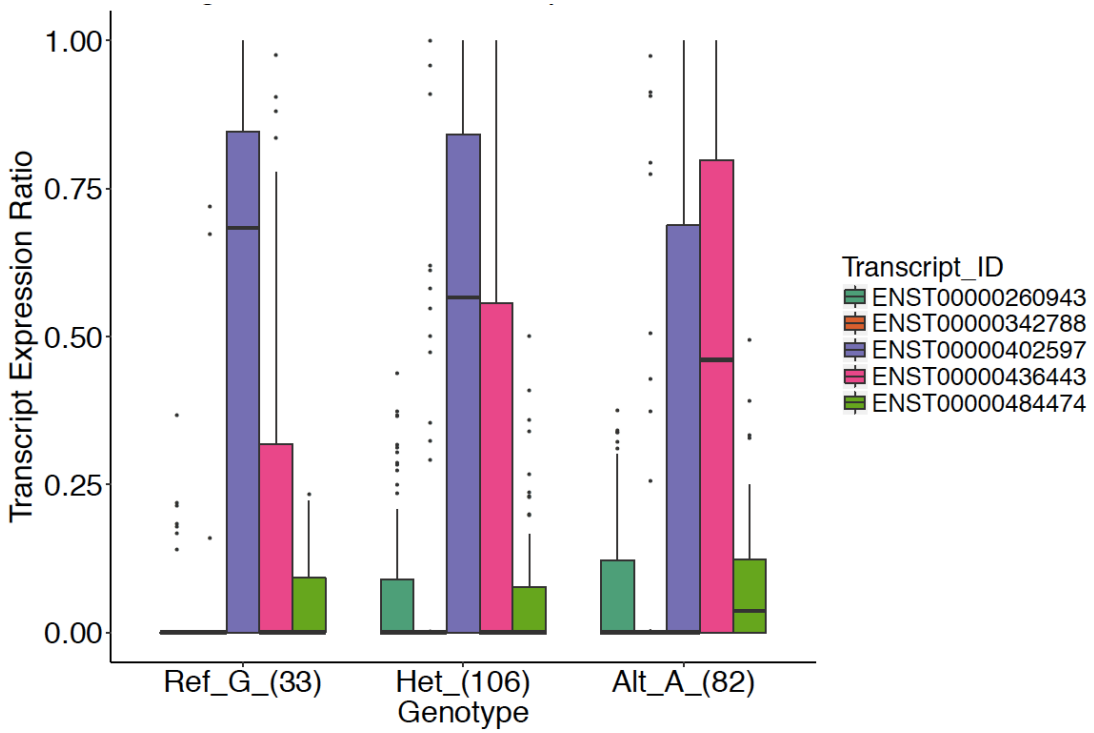


Figure 5.13 Change in ERBB4 transcript expression ratio. *Reference and alternative alleles are shown, and numbers of individuals of each genotype are detailed in brackets.*

An sQTL was identified for *CASP3* by sQTLseekeR linking the presence of a 1bp insertion caused by the rare SNP rs200872527 (MAF in 1000 Genomes 0.02) to a 29% decrease in the transcript expression ratio of ENST00000308394, and an equivalent increase of ENST00000523916 which skips the second exon of the gene (Figure 5.14). Leafcutter also identified an sQTL for *CASP3*, which related to inclusion of an intron with coordinates of 4:184638468-184649395, which exactly correspond to the first intron of *CASP3* (Figure 5.15). A sashimi plot generated by LeafViz shows that individuals with the alternative genotype (a genotype dosage of 2 as opposed to 0) for rs200872527 have an average of 72% of reads supporting the skipping of exon 2 via inclusion of an intron spanning exon 1 to exon 3, compared to only 37% of reads supporting this same transcript in individuals homozygous for the reference allele (Figure 5.16). Additionally, greater proportional utilisation of exon 1 compared to exon 2 can be seen in the reads aligned to CASP3 from two samples representative of the 0 and 2 genotype dosages (Figure 5.17). The Leafcutter sQTL was most significantly associated with rs4647609, only 4.76kbp upstream of rs200872527, implying that both packages detected the same sQTL event in this gene.

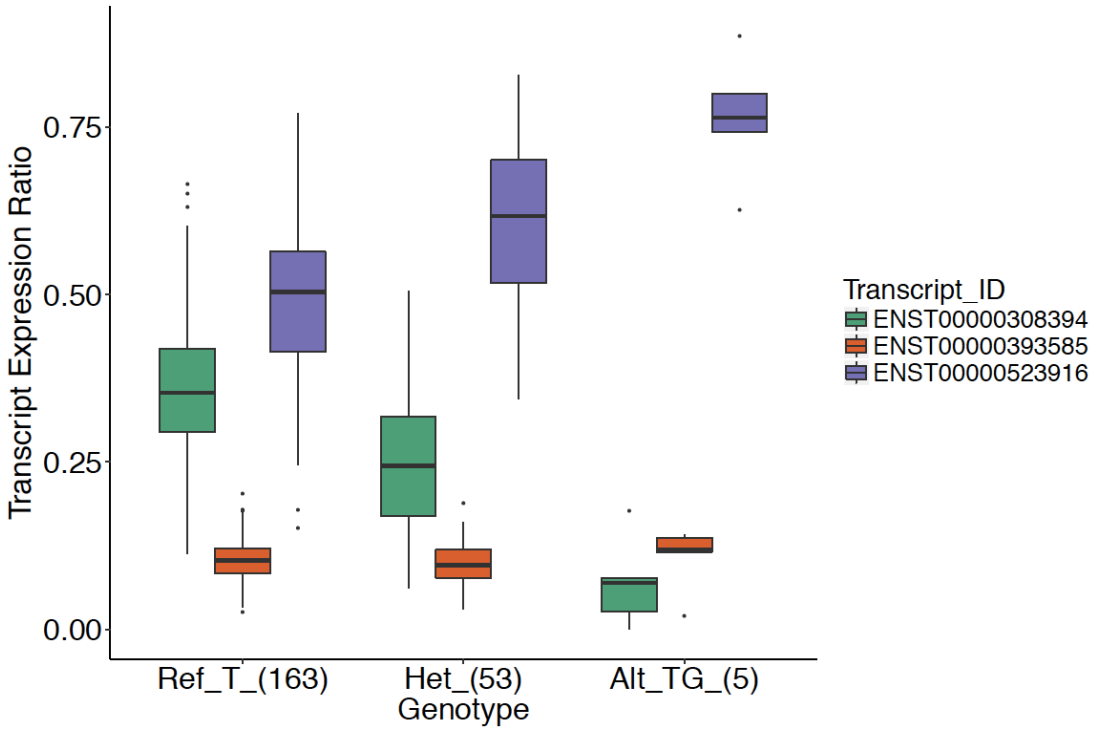


Figure 5.14 Change in CASP3 transcript expression ratio. *Reference and alternative alleles are shown, and numbers of individuals of each genotype are detailed in brackets.*

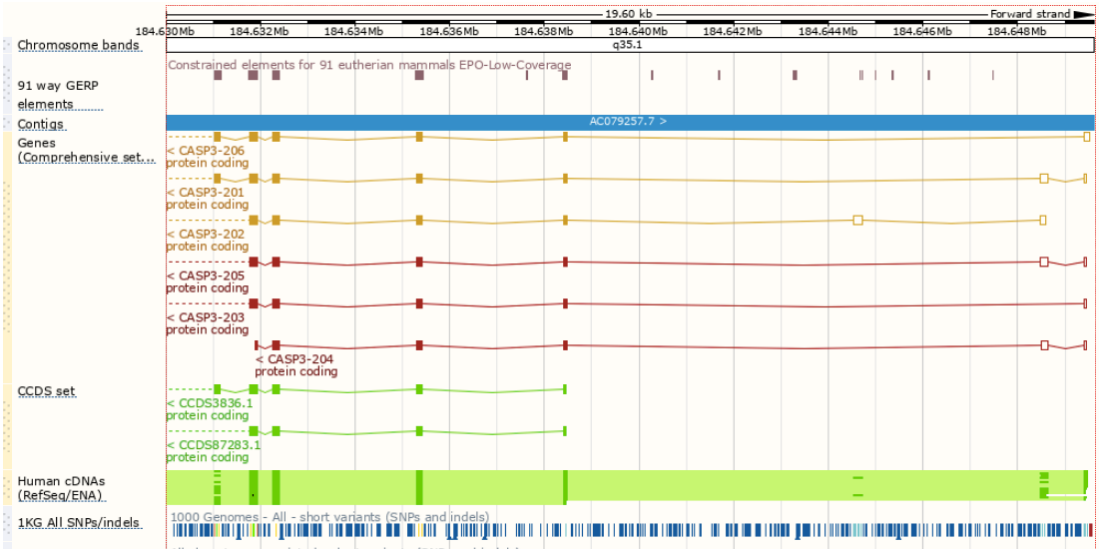
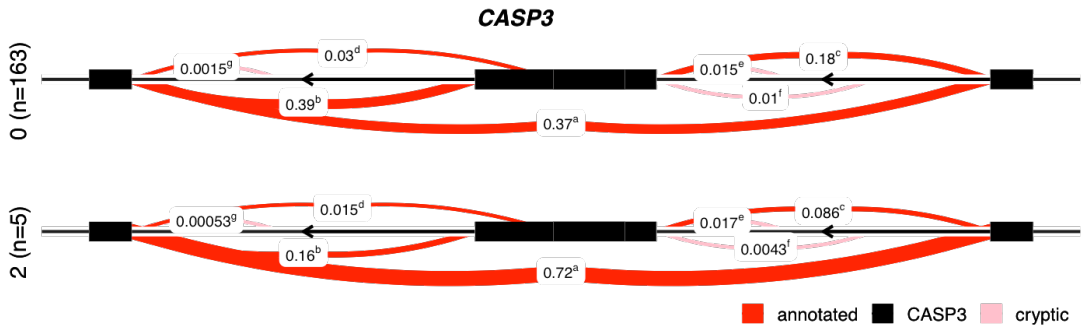


Figure 5.15 Ensembl browser view detailing the final intron of CASP3.
The intron with coordinates 4:184638468-184649395 corresponds to the first intron of transcript ENST00000523916 (CASP3-206) which skips the second exon of ENST00000308394 (CASP3-201)[328]*. Note, CASP3 is located on the negative strand.*



	chr	start	end	verdict	dPSI
a	chr4	184638468	184649395	annotated	0.365
b	chr4	184638468	184648465	annotated	−0.222
c	chr4	184648631	184649395	annotated	−0.105
d	chr4	184638468	184648503	annotated	−0.019
e	chr4	184648631	184649108	cryptic_fiveprime	−0.011
f	chr4	184648631	184649165	cryptic_fiveprime	−0.007
g	chr4	184638468	184639595	cryptic_fiveprime	−0.001

Figure 5.16 Sashimi plot detailing the changes in intron usage between individuals of genotype dosage 0 or 2 across the first 3 exons of CASP3.
Introns inferred by Leafcutter are labelled a-g. dPSI is mean change in percent spliced in.

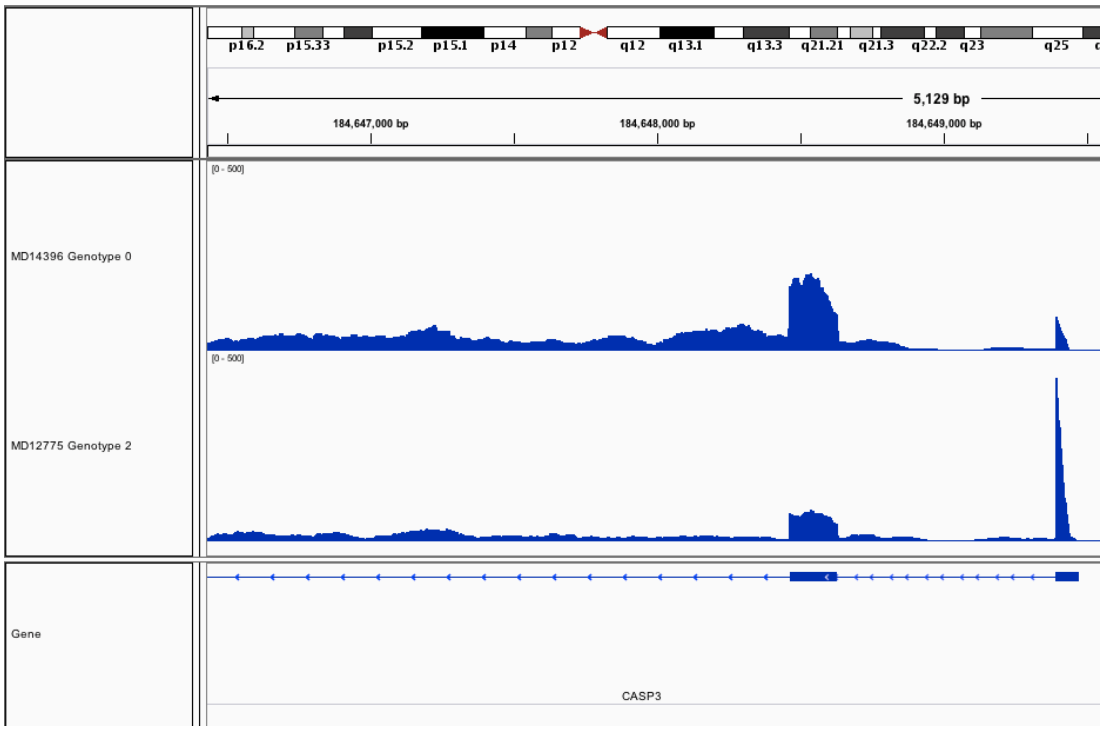


Figure 5.17 IGV screenshot detailing reads mapping to exons 1 and 2 of CASP3.
Bigwig files representing the density of reads mapped to CASP3 are shown for two separate samples which represent each of the opposing genotypes. It can be seen that the ratio of reads aligned to exon 1 compared to exon 2 is much greater in the individual of genotype 2, which promotes skipping of exon 2.

Leafcutter identified an sQTL in *POLE* corresponding to a change in the PSI of an intron with coordinates 12:132676184-132676546. These positions perfectly

correspond to intron 9 of transcript ENST00000537064, which is identical to the canonical protein-coding transcript ENST00000535270 until the presence of this intron, which introduces a nonsense-mediated decay motif[328] (Figure 5.18).

Figure 5.18 Ensembl browser view detailing intron 9 of POLE.
The presence of the intron with coordinates 12:132676184-132676546 causes a change from the protein-coding transcript ENST00000535270 (POLE-207) to the nonsense mediated decay transcript ENST00000537064 (POLE-210)[328].

5.5 Discussion

5.5.1 Variant effect prediction

This study found sQTLs to be more enriched in putatively functional exonic regions than non-exonic (Figure 5.2). Other groups have made the same observation: sQTLs identified from brain tissue by Takata *et al.* were enriched in exonic regions with an odds ratio of 3.84 compared to a background set of 48,068 non-significant, LD-pruned sQTL SNPs[264]. The Leafcutter authors found a greater proportion of exonic SNPs in the sQTLs they derived from GEUVADIS LCL cell lines compared to a background set of 200,000 randomly chosen variants from within the vicinity of genes[260].

The strongest enrichments of sQTLs in this study was within the SnpEff consequence class of splicing-related variants (Figure 5.2). This result was mirrored by many other groups investigating sQTLs: The GTEx Consortium, Takata *et al.,* the Leafcutter authors, Altrans authors and Humphrey *et al.* all found enrichments in splicing-related variants[200,258,260,264,265]. Although these classifications were relatively rare (0.34-1.48% of tested sQTL SNPs, Table 5.8), they produced the greatest enrichments relative to the background sets (Figure 5.2). This is why all significant

sQTLs were included in this facet of the analysis, not just the lead SNPs, in order to capture all possible causative SNPs within LD of an event. It could be that these SNPs with putative effects on splicing are the most likely to be responsible for the sQTLs, even if they are not the most statistically significant "lead" SNP. Where numbers were available, this pattern is seen in other studies: Takata *et al.* found only 0.3% of sQTLs to be in canonical splice sites and 0.6% in splice regions; however they represented some of the greatest enrichments relative to their non-significant background set with Z-scores of >9.0 and >4.0 respectively[264]. Only 0.44% of sQTLs identified by the sQTLseekeR authors from GEAUVADIS LCLs were within splice sites, but this was ~5x greater than the 0.09% of non-sQTL SNPs which were located in positions with the same classification[247].

There was significant enrichment of synonymous variants and missense variants in the sQTLs identified by this study compared to background sets (Figure 5.2), as was also seen by GTEx, Takata *et al.*, the Leafcutter authors and the Altrans authors[200,258,260,264]. Synonymous variants have been observed to affect splicing, and are theorised to mediate such effects by disrupting exonic splice enhancers and exonic splice suppressors (ESEs and ESSs)[157,403]. Hurst and Batada saw 17% reduced rates of somatic synonymous mutations in exon flanks, and a further lower density of somatic mutations in ESE motifs specifically, in cancer, implying a degree of purifying selection acting against mutations disrupting splicing in this way[442]. Synonymous variants have also been observed to be under negative selection in the germline of human populations[443,444] - implying that any aberrations caused by sQTLs could be significant in influencing predisposition to disease.

A greater significant enrichment of sQTLs within 5' UTR than 3' UTR regions was observed in this study, in agreement with Takata *et al.*, the Leafcutter authors and the Altrans authors[258,260,264]. This implies that regions at the beginning of a transcript may have more influence on its subsequent splicing than those at the end. Whilst the 3' UTR can affect transcript stability through the presence of micro-RNA binding sites[445], the 5' end can contain sequences influencing transcription initiation sites and mRNA capping[446]. Hurst and Batada also identified more purifying selection at 5' exon flanks than 3': further evidence that sequences earlier in a transcript may play a more critical role in influencing splicing[442]. The GTEx pilot study of nine tissues in 2015 found the opposite pattern for sQTLs identified by sQTLseekeR[200], however they only identified an average of 250 genes with sQTLs per tissue,

therefore they may not have identified the entire population of possible sQTLs. For the sQTLs they identified via Altrans (of which there was an average of 1,900 genes with sQTLs per tissue), there was again a greater enrichment within 5' than 3' UTRs[200].

These enrichments of the sQTLs identified in this study within functionally relevant sequences relative to randomly chosen background SNPs increases confidence in their validity. Even though the exonic and functionally-relevant SNP classifications are enriched within the vicinity of genes, and the sQTLs were specifically sampled from search windows containing genes, the background sets were drawn from windows matching the test sets, thus accounting for this bias.

Any variants associated with sQTLs by this study are likely to be relatively common in the wider population, given that thresholds were set requiring at least 5 individuals of each genotype group to be present in the cohort of 221 samples, meaning that they are less likely to be highly deleterious. This makes the enrichment of sQTLs in high effect consequence classes particularly noteworthy, and these are strong candidates for further functional investigation and fine-mapping.

5.5.2 Epigenetic and functional marks

When investigating sQTLs derived from dorsolateral prefrontal cortex brain tissue, Takata *et al.* identified a significant enrichment within ChIP-seq peaks for H3K4me3[264], a relationship which was also observed in this study (Table 5.9 and Figure 5.3). H3K4me3 was also one of the chromatin marks found to be most strongly enriched for sQTLs identified by the GTEx 2015 pilot study[200], and for sQTLs identified by Altrans from GEAUVADIS samples[258]. This makes functional sense as H3K4me3 is associated with TSSs of genes undergoing active transcription[419,447].

Surprisingly, Takata *et al.* calculated a significant depletion of sQTLs within H3K4me1 and H3K27ac - two other marks usually associated with active genes - whereas this study found an enrichment. The GTEx 2015 study also failed to find a significant enrichment of sQTLs identified by sQTLseekeR in H3K4me1 ChIP-seq peaks; however they did find an enrichment in H3K27ac[200]. H3K4me1 localizes around active enhancers[448] and can recruit the chromatin remodelling enzyme CHD7[449]. It is not as strongly enriched as H3K4me3 in 5' active promoter regions[419], which could explain its lower Z-score of 13.20 compared to 17.35 when permutated

against sQTL LD blocks. The search windows for sQTLs around genes would more likely encompass proximal promoter rather than distal enhancer regions. H3K27ac marks transcriptionally active regions[450] and primarily decompacts chromatin as a result of bringing a net neutral charge to histones, and by being antagonistic to the strongly repressive H3K27me3 mark, as a lysine cannot be simultaneously acetylated and methylated[451]. H3K27ac marks are often found to co-occur with H3K4me3[423], which could explain the very similar Z-scores of 17.14 and 17.35.

The discrepancies in H3K4me1 and H3K27ac enrichment in relation to the Takata analysis could be that they used a consensus set of histone marks derived from cell lines analysed by ENCODE, whereas this study used tissue-matched colonic mucosa epigenetic data from the Roadmap Epigenetics Consortium. Other studies have explored the tissue-specificity of epigenetic marks: Heintzman *et al.* found a Pearson correlation of 0.71 between H3K4me1, H3K4me3 and H3K27ac at promoters of 5 different cell lines, whereas they observed highly cell-type-specific histone modification patterns at previously identified and newly-predicted enhancer sequences[450]. The FANTOM Consortium produced an atlas of cell-type-specific enhancer sequences based on correlations between gene expression data, DNase I hypersensitivity, H3K4me1 and H3K27ac marks, and found many differences even between physiologically closely related cells, e.g. CD4+, CD8+, CD14+, CD19+ and CD56+ white blood cells[452].

This study saw a significant enrichment of sQTLs within H3K9ac and H3K36me3 marks (Table 5.9 and Figure 5.3), observations which were capitulated in GTEx data[200] and by the Altrans authors[258]. In its acetylated form, H3K9 is a marker of an active promoter state, and is localised to active promoters and enhancers[421]. H3K36me3 is deposited in the gene bodies of actively transcribed genes[424], and may serve to recruit de-acetylating enzymes as part of a negative feedback loop to prevent active genes in open chromatin undergoing aberrant intragenic transcription initiation in the presence of RNA Polymerase II[453,454]. H3K36me3 has also been hypothesised to play a role in facilitating alternative splicing because H3K36me3 peaks are highly enriched within nucleosomes specifically occupying exonic sequences[455], and the mark is hypothesised to recruit additional splicing factors to aid the excision of intronic sequences from around exon boundaries[184]. Therefore enrichment of sQTLs in both of these marks also makes functional sense.

This study showed a nominally significant depletion of sQTLs in H3K9me3 marks and no significant association with H3K27me3 (Table 5.9 and Figure 5.3). The GTEx consortium found a significant depletion of sQTLs in H3K27me3 regions, though they did not assess H3K9me3[200]. Both of these modifications mark inactive regions. H3K9me3 has been found to be enriched within 10kbp of the promoters of transcriptionally silenced genes[419], and it has also been shown to recruit DNA methyltransferases which subsequently promote the formation of condensed heterochromatin[422]. Tri-methylated H3K27 acts antagonistically to its activating, acetylated form[450], and is found in higher levels at silenced than transcriptionally active promoters[419]. Therefore depletion of sQTLs in these areas of repressed expression would be expected - and even though the associations did not pass multiple testing significance, there was no enrichment of sQTLs within these chromatin marks. The lack of significance in relation to overlap between sQTLs and either H3K9me3 or H3K27me3 could be because these marks covered the smallest percentages of the genome of all the histone modifications analysed; 12.1 and 20.9 Mbp respectively (Table 5.5). This means the random permutations used to ascribe significance will have been constrained towards fewer overlaps, producing a distribution that was less tractable for the actual numbers of overlaps with sQTLs to be significantly different from, even though they produced the fewest overlaps (73/965 and 91/965, Table 5.9). The relative paucity of H3K27me3 could be because it is especially implicated in the repression of genes during development, being laid down by a component of the PRC2 complex which is a key transcriptional repressor during the embryonic life stage of metazoans[456], whereas this epigenetic data was sampled from adult, differentiated colonic mucosa. Perhaps GTEx were able to find an enrichment because they profiled a wide range of tissues, and therefore may have been able to pick up signals from a more diverse set of tissue-specific promoters which are still active after differentiation.

Candidate regulatory elements were constructed by ENCODE from combined signals of the highest quality DNase I hypersensitivity, H3K4me3, H3K27ac, and CTCF ChIP-seq peaks[427]. There was a significant enrichment of sQTLs within these regions as well as the consensus DNase I hypersensitivity peaks combined across 125 cell lines profiled by ENCODE. Given these two region sets encompassed a similar proportion of the genome, it is an encouraging result that both were enriched for sQTLs with Z-scores of similar magnitudes of 8.19 and 9.35 respectively (Table 5.9). However, enrichment might have been expected given that these two regions

cover the greatest portion of the genome (548 and 387 Mbp respectively, Table 5.5), and the observed number of overlaps between them and sQTLs were high (849/965 and 855/965). The Altrans authors also observed a significant enrichment of sQTLs within DNase I hypersensitivity peaks[258].

sQTLs were found to be enriched in all of the ChromHMM states relating to transcription (TssAFlnk, TxFlnk, Tx, TxWk), except for those indicating bivalent transcription start sites (TssBiv, BivFlnk, Table 5.10). Similarly, there was enrichment within enhancer and genic enhancer regions (Enh, EnhG) but not bivalent enhancers (EnhBiv, Table 5.10). Bivalent regions are identified by a combination of both activating and repressive chromatin marks, often H3K4me3 within a broader region of H3K27me3[425,457,458], and are theorised to be "poised" ready to release repression from transcription factors which promote differentiation[420,456]. Therefore, similar to the lack of enrichment of sQTLs within the individual H3K27me3 mark, it could be that because these analyses were performed on adult, differentiated tissue, there were fewer regulatory regions being held in this bivalent state. Raj *et al.* found a similar enrichment of sQTLs in actively transcribed but not bivalent regions when analysing expression data from dorsolateral prefrontal cortex brain tissue[265]. This was mirrored by the Leafcutter 2018 study which observed an enrichment of sQTLs from GEAVADIS CEU LCLs in strong enhancers and active promoters, but not poised promoters[260]. The Roadmap Epigenetic Consortium found enrichment of evolutionary conservation of non-coding sequences identified as containing enhancers and promoters[425,459], indicating the potential disease-relevance of germline variants in these regions.

This study observed a significant depletion of sQTLs in regions assigned a chromatin state of quiescent (Quies, Table 5.10), as would be expected due to these mainly representing intergenic regions[458]. However, there was no depletion in regions predicted to have high levels of polycomb-protein binding which cause compaction of chromatin and result in gene silencing[460] (ReprPC or ReprPCWk). Similarly, depletion within heterochromatic regions (Het), which usually correlate with lower gene expression[461], was not significant. The Leafcutter authors did observe a significant depletion of GEAUVADIS sQTLs in regions assigned as being heterochromatic and polycomb-repressed[260], and Raj *et al.* found a significant depletion within polycomb-repressed and quiescent regions, and a nominally significant depletion in heterochromatic regions[265]. The lack of significant depletion

of sQTLs in heterochromatin in this study is a surprising result given that there was a significant enrichment of sQTLs within regions predicted to contain zinc finger repeat sequences (ZNF_Rpts), and that heterochromatin is often targeted to repetitive regions such as these to prevent them undergoing aberrant homologous recombination[462]. Perhaps this discrepancy between heterochromatic and zinc-finger domains could be explained by the regions assigned as ZNF_Rpts covering a smaller overall region of the genome than heterochromatin (7x fewer bases covered (3.5Mbp vs 25.1Mbp) and 5.7x fewer regions (2,281 vs 12,987), Table 5.6). The Roadmap Epigenetics Consortium ran ChromHMM predictions with outputs ranging from 10 to 25 states, and chose to use 15 different states because larger numbers failed to sufficiently distinguish between biologically distinct regions[425]. However perhaps some of the resulting states became too rare for this kind of enrichment testing to be appropriate: Takata *et al.* didn't test for enrichment of sQTLs in any ChIP-seq dataset which had fewer than 50,000 peaks[264], which the majority of the ChromHMM states do not have (Table 5.6). Therefore a limitation of analysing individual ChromHMM states could be the relative scarcity of the rarer classes. A new implementation of ChromHMM trained on continuous annotation which accounts for the strength of experimentally-derived peaks rather than using discrete presence or absence calls of features may bring additional certainty to the study of integrated epigenetic states[463].

A caveat to all the observed enrichments of sQTLs in relation to epigenetic marks is that the search windows for sQTLs were constrained around genes or inferred introns, and many of the chromatin marks analysed tend to be located in the proximity of genes - regardless of the state of their activity. This co-localization may have served to intrinsically augment the strength of all enrichments, which could explain why there were nominally significant or non-significant depletions of sQTLs in repressive H3K9me3 and H3K27me3 regions when significant depletions would have been expected. However, the local Z-score plots do tend to show distinct peaks and troughs where expected more often than not, which indicates that these observations are not purely artefactual and that the specific position of the test set of regions relative to the permuted set does influence the significance of enrichments.

The colonic tissue sample E075 from which the Roadmap Epigenetics Consortium derived the individual ChIP-seq peaks used in this chapter was from a female donor. However, network analysis detailed in the first results chapter concluded that there

was little difference between the gene expression networks in colonic mucosa of males and females, which would imply that epigenetic regulatory profiles obtained by ChIP-seq should be applicable between genders. Tissue specificity was the most important consideration for this analysis of alternative splicing, hence sample E075 was used for the majority of ChIP-seq peaks as opposed to a conglomeration across multiple samples.

5.5.3 Comparing eQTLs and sQTLs

This study found that only a small percentage (2.14-2.79%, Table 5.11) of lead sQTL SNPs are also lead eQTLs. Raj *et al.* saw a similarly low degree of sharing noting that only 42 of 9,045 lead sQTL SNPs from DLPFC were also lead eQTL SNPs (0.46%)[265]. From their 2016 analysis, the Leafcutter authors found that of 275 genes for which there was an eQTL and an sQTL, only 14 variants were the lead SNP for both QTL types, and that >74% of lead sQTL SNPs did not have any detectable effect on total expression levels of the genes they affected[261].

Circular permutation found significant overlaps of LD blocks containing sQTLs with LD blocks containing eQTLs identified from the same Scottish dataset or from GTEx tissues (Table 5.12, Table 5.13). However, whilst there were significant enrichments compared to random circular permutations, when SNPs were assigned to the same set of LD blocks, the actual number of blocks which overlapped was low relative to the number of different sQTL and eQTL-containing blocks tested. 257 sQTL LD blocks overlapped with eQTL blocks from the Scottish dataset, out of 10,110 sQTL and 11,478 eQTL LD blocks respectively (Table 5.12)

Of 4,148 genes for which there was an sQTL identified by either sQTLseekeR or Leafcutter, 68.8% also had an eQTL identified in the Scottish cohort (Table 5.11). When identifying eQTLs and sQTLs from nine separate tissues, the GTEx consortium observed a similar degree of co-occurrence, with up to 70% of genes for which sQTLseekeR identified an sQTL also having an eQTL (13%-70%, mean 48%)[200]. This level of agreement is likely attributable to the fact that the level of expression influences the propensity for a gene to have either an eQTL or an sQTL associated with it.

Therefore it appears that although many of the same genes possess both an sQTL and an eQTL, the signals often originate from independent SNPs located in separate LD blocks. This is an important observation of the apparent independence

between the majority of sQTL and eQTL events identified from colonic mucosa, and implies that accounting for loci implicated in alternative splicing phenotypes may add additional power when prioritising disease-linked GWAS variants for colorectal cancer via colocalisation analyses or TWAS[265,401,464–466].

5.5.4 GWAS enrichment

Yang *et al*. demonstrated that in the absence of population structure, the lambda genomic inflation factor is dependent on heritability of the quantitative trait in question, the number of causative variants for the trait, the sample size analysed and the LD structure within the cohort[436]. Both the mean and median lambda inflation increase with heritability, and the median increases as a function of the number of causative variants, because more SNPs will then be in LD with these variants and so their p-values of association will deviate further from the null expected chi-squared distribution. Yang *et al.* observed mean and median lambda inflations of 1.035 and 1.029 respectively in a GWAS for height using 294,831 genotyped SNPs in a relatively small cohort of 3,925 unrelated individuals in which they had previously determined there was negligible detectable population structure. This indicates that some degree of inflation is detectable even under such theoretically ideal conditions, and therefore the background genomic inflation for a given cohort needs to be accounted for. By sampling 100,000 randomly selected loci, they estimated there to be an average of 188 SNPs in LD at each locus, with an average r-squared of 0.026. In a larger meta-GWAS cohort of 133,000 individuals and 2.8M genotyped and imputed variants, Yang *et al.* observed mean and median lambda inflations of 1.95 and 1.55 respectively in the p-values of association with height. The meta-GWAS of 58,640 individuals from which CRC association p-values were utilised for this analysis presented mean and median lambda inflation values intermediate between these two cohorts studied by Yang *et al.*, of 1.243 and 1.236 for sQTLseekeR and 1.278 and 1.274 for Leafcutter (Figure 5.9), as would be expected given its size and the relatively lower estimates of heritability of CRC of between 7.42-26.0% compared to 54% estimated for height. The fact that the median lambda inflation of CRC meta-GWAS p-values corresponding to thresholded Leafcutter sQTL SNPs (1.595) deviated significantly from a corresponding null distribution of MAF and window-matched SNPs (p-value=0.00273) implies that the Leafcutter sQTLs are more commonly in higher LD with causative SNPs for CRC than would be expected by chance, and that they have more significant p-values of

association with CRC predisposition than would be expected by chance (Figure 5.9). The fact that the median lambda inflation (1.305) of the thresholded sQTLseekeR sQTL SNPs was not significantly different from its null distribution (p-value = 0.682) could be as a result of it not being as prolific as Leafcutter at identifying sQTLs. In addition, the meta-GWAS of CRC from which summary statistics of association were used in this analysis has since been superseded by larger meta-analyses which have uncovered dozens more loci significantly associated with CRC predisposition[96,231], meaning that it is possible this analysis underestimated the significance of sQTL SNPs, and the sQTLs identified by sQTLseekeR may constitute a significantly associated set if tested in the context of these updated association statistics.

The Leafcutter authors observed significant enrichment of sQTLs they identified from two different GEAUVDIS populations of LCLs in GWAS association statistics for rheumatoid arthritis and multiple sclerosis, and observed clear deviation from the expected quantiles in QQplots[260,261]. The observed deviation was less marked in this study, however they did not filter their sQTLs for effect size or for read counts supporting the inferred introns, and they compared against all genome-wide SNPs not just those within the search window of the algorithm, meaning their inflations may have been exaggerated[260].

Takata *et al.* found enrichment of PSI sQTL SNPs identified from DLPFC tissue in loci associated with schizophrenia using a 1-talied Fisher's test (OR=3.72, p-value=9.90E-05). They defined disease-associated loci as any SNPs in LD of r-squared >0.6 with lead GWAS SNPs for schizophrenia sourced from the NHGRI GWAS catalog[429] or the Psychiatric Genomics Consortium[467]. They defined a set of MAF-matched (by 0.02 bins) non-sQTL SNPs as those which had nominal p-values >0.05, and used these in conjunction with the significant sQTLs to construct a 2x2 table with columns: sQTL SNPs vs non-sQTL SNPs and rows: within vs outwith disease-associate loci, from which to perform their Fisher's test.

5.5.5 sQTLs in cancer-relevant genes

Hereditary mixed polyposis syndrome (HMPS) is a rare Mendelian trait which predisposes to the presence of adenomas and hyperplastic polyps, with a mean age of first presentation of 33 years[468]. The causative genomic aberration has been traced by multiple groups to germline copy number duplications of a 20kbp region

encompassing exons 3 to 6 at the 3' end of the *SCG5* gene and extending into the region upstream of the BMP agonist *GREM1*[469,470]. *SCG5* is a secretogratin gene which binds cargo proteins in the endoplasmic reticulum and chaperones them through the secretory pathway to the cell surface[471]. An sQTL was found in this study which lead to a 44% increase in relative expression of a transcript of *SCG5* which includes an extra alanine at the beginning of exon 3. The variant most significantly associated with this event, rs72715244, is located 53kbp upstream of rs4779584 which has been implicated in CRC predisposition by a meta-GWAS of 11,769 cases and 14,328 controls[472].

PTPRT is a member of a super family of tyrosine phosphatases which can act as tumour suppressors by tempering kinase signalling cascades mediating cell-cell adhesion[473]. *PTPRT* is the family member which most commonly suffers deleterious mutations in CRC[474], and has also been observed to undergo promoter hypermethylation as a mechanism of being downregulated[475]. This study identified rs6030443 as being the SNP most significantly associated with an alternative splicing event causing a 29% increase in expression of a non-canonical transcript of *PTPRT* which does not possess the last 7 exons of the gene. These exons include regions coding for the intracellular phosphatase domains[476]. It is possible that reduced expression of the full length version of this tumour suppressor gene in the colonic mucosa throughout the course of an individual's lifespan may increase their likelihood of developing CRC.

ERBB4 (previously known as *HER4*) is a member of the subfamily of EGFR receptor tyrosine kinases, which promote cell proliferation, differentiation and migration upon binding of ligands including EGF[477]. Deep sequencing of 91 hotspot regions of 653 cases of sporadic CRC by Malapelle *et al*. determined that *ERBB4* was mutated in 0.6% of cases[440], and the presence of such mutations stratified survival in 276 TCGA patients, corresponding to lower mean survival times (p-value=0.00942)[441]. RNA-seq analysis of 250 CRC samples from the Vanderbilt Medical Centre found *ERBB4* to be over expressed relative to normal tissue in samples from all of the CRC stages I-IV, and its over expression in mouse xenografts harbouring *APC^{min}* and v-Ha-Ras produced tumours of double the size compared to those with wildtype expression levels[441]. This study found an sQTL event comprising a 27% increase in expression of an *ERBB4* transcript with an additional final exon in relation to rs10192485, which is located 146kbp upstream of another intronic *ERBB4* variant,

rs10932384, which has been significantly associated with recurrence and overall survival in renal clear cell carcinoma[478].

CASP3 is part of the caspase family of proteases involved in the execution-phase of cellular apoptosis, which are cleaved to form two subunits which then dimerize to constitute the active version of the enzyme[479]. Caspases often undergo mutations in cancer which allow tumour cells to evade programmed cell death, with higher levels of cleaved CASP3 predicting better prognosis and survival in CRC[480]. Both sQTLseekeR and Leafcutter appeared to identify the same sQTL for *CASP3*, which corresponded to greater relative expression of a transcript which skipped the second exon of the gene. It could be that perturbing the sequence of such a key tumour suppressor could reduce the ability of cells to respond to pro-oncogenic signals and curtail proliferation of an incipient cancer cell by instructing them to undergo programmed cell death.

POLE has strong links to CRC, both in terms of germline predisposition[481] and somatic mutations which lead to hypermutator phenotypes[482]. In this study, Leafcutter identified a variant, rs4077170, which was associated with inclusion of an intron causing nonsense mediated decay of the canonical protein-coding *POLE* transcript. POLE is the DNA polymerase most commonly used for synthesising the leading strand[483]. The mean counts supporting inclusion of the nonsense mediated decay intron across all 3 genotype groups was relatively low, 73.9, and the allele of the variant associated with inclusion of the intron had a frequency of 0.70 in the 1000 Genomes EUR population (Table 5.17). This means that the sQTL itself is likely neutral, as opposed to deleterious.

Chapter 6 Conclusions

6.1 Results

This project is the first to comprehensively identify *cis*-sQTLs in colonic mucosa, by applying two different algorithms with complementary approaches; the transcript-aware sQTLseekeR[247] and intron-centric Leafcutter[261], to RNA-seq from a Scottish cohort of 221 genotyped individuals. The variants associated with these sQTL events tended to fall within gene bodies, or lie proximal to the 5' and 3' ends of genes. Thresholds were implemented to prioritise a set of higher confidence sQTLs with larger effect sizes originating from features with greater expression. These retained the top 7.79% of sQTLseekeR events and 8.37% of Leafcutter events.

Compared to background sets of non-significant SNPs matched for MAF and position, sQTL SNPs were most highly enriched within splicing related variants. There were also enrichments within synonymous coding variants, and there was a greater enrichment within 5' than 3' UTRs, perhaps indicating that variants situated earlier in a transcript can have a greater influence on splicing. Using circular permutations, the LD blocks containing sQTLs were found to be enriched within a range of histone modifications indicating actively transcribed genes, according to data obtained by the Roadmap Epigenetics Consortium from colonic mucosa tissue[418]. The sQTLs were similarly enriched within active chromatin states and significantly depleted in quiescent regions.

68% of genes for which an sQTL was identified in this study also possessed an eQTL according to the same expression data; however only 2.79% of lead sQTL SNPs were also lead eQTL SNPs, and only 2.54% of LD blocks containing sQTLs overlapped with LD blocks containing eQTLs. This suggests that there is independence between these two classes of QTLs, and therefore this study expands our knowledge of non-coding variants influencing transcriptional regulation in human colonic mucosa.

Thresholded sQTL SNPs were tested for genomic inflation in relation to loci predisposing to CRC using their corresponding p-values from a meta-GWAS of 20,181 cases and 30,822 controls. The median lambda inflation for the thresholded Leafcutter sQTL variants deviated significantly from their corresponding null distribution, while the thresholded sQTLseekeR SNPs did not. This could be

because there were not as many SNPs in the sQTLseekR set, or because the meta-GWAS from which GWAS p-values were obtained may not have represented a comprehensive set and as a result may have underestimated the significance of associations for certain loci.

sQTLseekeR and Leafcutter together identified sQTLs in 9 genes with associations to CRC listed in the NHGRI-EBI GWAS catalog, 4 genes curated in the COSMIC database as being relevant to CRC progression, and a further 29 genes implicated by COSMIC in any of the cancers they assess. *PTPRT* is the tyrosine phosphatase most commonly mutated in CRC[474], and sQTLseekeR identified an sQTL causing a reciprocal 29% change in expression of the canonical transcript and one which lacks the final 7 exons coding for intracellular phosphatase domains[476]. sQTLseekeR and Leafcutter both identified sQTLs affecting the relative expression of *CASP3* transcripts, a tumour suppressor gene involved in regulating programmed cell death, and which if not properly expressed could reduce the likelihood of abnormally proliferative cells being ablated from the colonic mucosa. Leafcutter identified an sQTL in *POLE*, which is strongly implicated in both predisposition to and progression of CRC[481,482], corresponding to inclusion of a non-canonical intron which produces a nonsense mediated decay transcript. The mean counts supporting the presence of the intron were relatively low, and the SNP most significantly associated with the event had a frequency of 0.70 in the 1000 Genomes EUR population, implying it represents a tolerable, neutral change.

6.1.1 Limitations

This analysis of 221 individuals is among the largest cohorts in the literature from which sQTLs have been identified. However, the requirement of ≥5 individuals possessing each of the three possible genotypes of a biallelic SNP means that this study is still limited to identifying relatively common variants tagging loci associated with an sQTL. There may be more rarer variants associated with sQTLs to be discovered, which could potentially be under stronger purifying selection in the population. Additionally, this study has been carried out in a Scottish cohort. This will have made population-specific signals clearer and more readily identifiable, however it limits the scope of any conclusions to a Scottish population.

The decision was taken to combine two batches of RNA sequencing for this analysis, given the findings of previous studies that using greater numbers of

samples leads to the identification of more sQTLs[246,247]. sQTLseekeR should have been immune to batch effects by using ratios of transcript expression as opposed to absolute values, and FastQTL was able to take as covariates principal components derived from the Leafcutter intron usage values which captured the batch effects.

Samples were also combined from both the left and right sides of the colon. There were not sufficient right-sided samples to perform an adequately powered separate sQTL analysis. It could be that samples originating from different sides of the colon may have had different transcripts preferentially expressed, which could have reduced power to identify sQTLs in this study. However, previous analyses of total gene expression found few differentially expressed genes between the two sides[47], implying their transcriptional divergence is low.

The identification of sQTLs by sQTLseekeR rests on the accurate quantification of transcript expression. The alignment-free quantification algorithm Salmon has been demonstrated to outperform traditional two-step quantification pipelines[301]. Despite this, all transcript quantification is inherently probabilistic. There are certain conditions under which the accuracy of transcript-level quantification has been called into question, such as in the assignment of TPM values to transcripts shorter than the read length of the library[356]. This discrepancy originates from the method used by alignment-free algorithms to assign effective transcript length to features, however this analysis should be mostly immune to such shortcomings because raw counts instead of TPMs were supplied to sQTLseekeR.

Recent advances in the field of long-read technologies by PacBio and Oxford Nanopore facilitate the capture of full-length transcripts, and so can definitively survey the transcriptome without requiring probabilistic inference of transcript structure[484]. Such technologies are still currently less cost effective than Illumina short-read sequencing, and suffer from a lower per-base accuracy; however future improvements may see them become more widely adopted. They currently find application in hybrid approaches with short-read technologies for the assembly of novel transcriptomes from species with little existing sequencing data by providing scaffolds to which the shorter reads can be aligned to correct any sequencing errors[485].

A *de novo* transcriptome assembly was not performed in this study because the sequencing was from normal tissue samples which would be assumed to possess a

fairly regular transcriptome, well represented by the Ensembl gene build v88[328]. Gupta *et al.* performed full length isoform sequencing of single cells from a variety of mouse cerebral tissues and identified a multitude of putatively novel transcripts not captured by the latest gene builds[486], however they were often rare and may represent transcriptional noise as opposed to new canonical transcripts. If sequencing was performed on CRC tumour samples then a *de novo* transcriptome assembly may be more relevant as there could be cancer-specific transcripts expressed.

None of the sQTLs in this study have yet been validated by an alternative quantification technology such as qPCR. The actual functional activity of the SNPs associated with the events could also be tested by minigene splicing reporter assays[487]. The sQTLs relating to genes associated with CRC predisposition or progression would be the highest priority candidates for follow up using such techniques.

6.2 Future work

6.2.1 Use of sQTLs for fine-mapping causative GWAS variants

sQTLs have been shown to significantly contribute to predisposition of other complex traits including Alzheimer's[265], schizophrenia[264], multiple sclerosis and rheumatoid arthritis[261], sometimes to a greater degree than eQTLs[261].

The sQTLs identified in this project could be used to perform colocalisation in association with GWAS summary statistics in order to fine-map the most likely causative variants in loci tagged as contributing to CRC predisposition. Nica *et al.* developed a regression framework testing the proportionality of coefficients for two traits (in this case GWAS association and eQTL coefficients) across a set of the same SNPs within a given locus to assess the likelihood of colocalisation[464]. Whilst this approach accounts for LD structure, a limitation is the set of SNPs to test must be defined *a priori* and genotype data at the resolution of each individual is required[464]. Giambartolomei *et al.* developed the coloc package which uses approximate Bayes factors to assess whether a shared causal variant is likely present in a region given the LD at the locus[488]. It is able to use simply summary GWAS and QTL association statistics, and only suffers a minor loss of power when using imputed variants. However it makes assumptions that the causal SNP is

present in the set of SNPs analysed from the locus, and assumes that there is only a single causal SNP present for each or both traits. Zhu *et al.* propose the summary data-based Mendelian randomization analysis (SMR) in concert with heterogeneity in dependent instruments (HEIDI)[466]. They perform Mendelian randomization (MR) for genes with eQTLs in the vicinity of a GWAS locus, checking whether the genotype groups corresponding to different expression levels also harbour more individuals of the given disease trait than would be expected by chance. In order to distinguish between pleiotropy, where the same SNP is causative of both the eQTL and the disease trait which would allow accurate prioritisation of disease genes, and linkage, whereby the causative eQTL SNP is simply in LD with the causative disease SNP and therefore the corresponding eQTL gene is not of relevance to the disease, they carry out MR for all SNPs in LD with the eQTL. If there is heterogeneity of the MR results then it implies simple linkage, however if there is no heterogeneity then it demonstrates consistency between the expression and disease phenotypes in each genotype group of the variant, meaning the variant likely influences both the eQTL and the disease trait, thus implicating the gene in the aetiology of the disease[466].

Law *et al.* performed a GWAS for CRC using 34,000 cases and 76,000 controls of European ancestry and were able to identify 31 new risk loci. They used SMR based on eQTLs identified from an in-house cohort and GTEx transverse colon samples, and were able to fine-map with high confidence the genes through which the associated variants at three of these loci mediated their effects, though almost 50% of loci had no *cis*-eQTLs detectable[96]. Whilst the majority of colocalisation studies to date have used eQTL SNPs, given the observation in this analysis of frequent independence between eQTLs and sQTLs from colonic mucosa, combining both sets of events together may allow greater probing of CRC-relevant GWAS signals.

Liu *et al.* warn against the now abundant availability of eQTLs confounding such colocalisation studies[489]. However the tissue-specificity of this study in terms of the source of QTLs and the associated GWAS trait, combined with the expression and effect size thresholding applied to the sQTLs, should have reduced the risk of such false positive associations being observed.

6.2.2 Transcriptome-wide association studies (TWAS) and sQTLs

Transcriptome-wide association studies have recently become adopted as a method of integrating variants known to influence gene expression regulation in order to increase the power of GWAS studies to identify new predisposition loci for complex traits. Li *et al.* used sQTLs and eQTLs they identified from LCLs using Leafcutter as input to the S-PrediXcan association algorithm[262] to identify new loci predisposing to rheumatoid arthritis[260]. They were able to identify 18 new associations using the PSI splicing phenotypes from Leafcutter, of which 13 were not able to be identified using gene expression alone[260]. Similarly, Raj *et al.* observed more associations when using splicing vs gene expression phenotypes from dorsolateral prefrontal cortex in two separate TWAS to search for novel associations with Alzheimer's, finding 16 vs 5 loci respectively, of which 8 were not able to be identified from GWAS alone[265]. The scope to further expand CRC association studies via TWAS using the sQTLs identified in this analysis is clear.

6.2.3 Polygenic risk scores

Whilst they may not provide information about the clinical prognosis of CRC, polygenic risk scores have proven useful in predicting the likelihood of individuals developing CRC[490]. If additional CRC risk loci could be discovered through the inclusion of sQTLs in TWAS analyses, their inclusion into such metrics would further increase the accuracy of prediction of risk. This is particularly relevant to CRC given that survival rates are highest if the cancer is identified and treated in its earlier stages[5,6]. Screening resources could be more effectively targeted to individuals at higher risk, and such monitoring will become increasing feasible at a large scale as the costs of genotyping and sequencing continue to fall.

6.2.4 Inclusion of other cohorts to increase power

Future work could include expanding the size of the cohort analysed, so as to be able to detect signals from rarer variants and to increase certainty in transcript expression quantification for each genotype group. RNA-seq data is available for sigmoid and transverse colon samples collected by the GTEx Consortium, which could be used to call sQTLs. However such samples were demonstrated to have some of the poorest RNA integrities following ischaemia[393,394], and the collection procedure led to significant amounts of stroma being captured along with the mucosa, which would dilute the tissue-specific signal from the precise tissue of

origin of CRC. In contrast, the samples used in this study specifically had stroma and muscle layers removed before sequencing. Analysis of sequencing from unrelated populations may also identify further sQTLs not detectable in this Scottish cohort.

6.2.5 Analysing CRC expression in relation to germline sQTLs

This investigation could be extended using RNA-seq from matched tumour samples from the same individuals from which normal tissue was obtained. It could be ascertained whether sQTLs observed in normal tissue persist into tumours, or whether patients who did not possess the germline variants go on to acquire them somatically. This could provide evidence of putative non-coding sQTL driver events, which have been hypothesised by other groups[276]. Data obtained in-house by Dr Debbie Baishnab using RNA Scope on histological samples of crypts from normal and tumour samples from the same individuals suggested that the SHROOM2 eQTL, which has been linked to CRC predisposition[98], does persist from normal into tumour tissue. Or alternatively it could be that sQTLs which predispose to CRC simply create a transcriptional environment which is favourable for the initiation of cancer, but do not themselves contribute to progression. This supposition would be supported by the fact that polygenic risk scores based on the current known loci associated with CRC did not prove effective in predicting survival outcomes in patients with CRC[245].

Having matched tumour RNA-seq available would allow the classification of cancers into one of the four consensus molecular subtypes of CRC[67]. It could then be tested whether certain germline sQTLs predispose to the development of cancers belonging to particular molecular subtypes, which have been shown to have different clinical outcomes and responses to treatments[491].

6.3 Summary

This project has performed a comprehensive analysis of alternative splicing quantitative trait loci in normal human colonic mucosa from a Scottish cohort of 221 individuals.

Given the significant lambda inflation of Leafcutter sQTLs within a meta-GWAS, the number of CRC-specific and cancer-relevant sQTL events identified by both

packages, and the precedent of other studies observing sQTLs to significantly predispose to other complex traits[260,264,265], it is likely that common germline sQTLs explain some portion of the non-coding variants which predispose to CRC.

Chapter 7 References

1. Cancer Research UK: Cancer incidence for common cancers. (2015). Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared.

2. Cancer Research UK: Cancer mortality for common cancers. (2016). Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality.

3. Cancer Research UK: Bowel cancer incidence and mortality statistics. (2014).

4. American Joint Committee on Cancer colorectal cancer staging guidelines. (2019). Available at: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html.

5. Benson, A. B. I. *et al.* American Society of Clinical Oncology Recommendations on Adjuvant Chemotherapy for Stage II Colon Cancer. (2004). doi:10.1200/JCO.2004.05.063

6. *Former Anglia Cancer Network, Five-Year Survival by Stage, Adults (aged 15-99 years), 2002-2006*. (2006).

7. Hewitson, P., Glasziou, P., Watson, E., Towler, B. & Irwig, L. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. *Am. J. Gastroenterol.* **103**, 1541–9 (2008).

8. *Public Health England: Routes to diagnosis 15 update - colorectal cancer*. (2015).

9. Desantis, C. E. *et al.* Cancer Treatment and Survivorship Statistics , 2014. (2014). doi:10.3322/caac.21235.

10. Lumley, J. S. P., Craven, J. L. & Aitken, J. T. *Essential Anatomy*. (Churchill Livingstone, 1995).

11. Large intestine diagram. *Wikimedia Commons* (2004). Available at: https://commons.wikimedia.org/wiki/File:Intestine.png.

12. Humphries, A. & Wright, N. A. Colonic crypt organization and tumorigenesis. *Nat.*

Rev. Cancer **8**, 415–424 (2008).

13. Kim, Y. S. & Ho, S. B. Intestinal Goblet Cells and Mucins in Health and Disease: Recent Insights and Progress. *Curr. Gastroenterol. Rep.* **12**, 319–330 (2010).

14. Li, H. J., Ray, S. K., Singh, N. K., Johnston, B. & Leiter, A. B. Basic helix-loop-helix transcription factors and enteroendocrine cell differentiation. *Diabetes, Obes. Metab.* **13**, 5–12 (2011).

15. McGuckin, M. A., Eri, R., Simms, L. A., Florin, T. H. J. & Radford-Smith, G. Intestinal barrier dysfunction in inflammatory bowel diseases. *Inflamm. Bowel Dis.* **15**, 100–113 (2009).

16. Gerbe, F. *et al.* Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J. Cell Biol.* **192**, 767–780 (2011).

17. Gerbe, F. *et al.* Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226–230 (2016).

18. McGrath, A. *Anatomy and Physiology of the Bowel and Urinary Systems*. (Blackwell, 2005).

19. American Cancer Society colorectal cancer staging. (2018). Available at: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html.

20. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nat. Rev. Mol. Cell Biol.* **15**, 19–33 (2014).

21. Valenta, T. *et al.* Wnt Ligands Secreted by Subepithelial Mesenchymal Cells Are Essential for the Survival of Intestinal Stem Cells and Gut Homeostasis. *Cell Rep.* **15**, 911–918 (2016).

22. Fevr, T., Robine, S., Louvard, D. & Huelsken, J. Wnt/beta-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol. Cell. Biol.* **27**, 7551–9 (2007).

23. Degirmenci, B., Valenta, T., Dimitrieva, S., Hausmann, G. & Basler, K. GLI1-expressing

mesenchymal cells form the essential Wnt-secreting niche for colon stem cells. *Nature* **558**, 449–453 (2018).

24. Kosinski, C. *et al.* Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 15418–23 (2007).

25. Beumer, J. *et al.* Enteroendocrine cells switch hormone expression along the crypt-to-villus BMP signalling gradient. *Nat. Cell Biol.* **20**, 909–916 (2018).

26. Lopez-Arribillaga, E. *et al.* Bmi1 regulates murine intestinal stem cell proliferation and self-renewal downstream of Notch. *Development* **142**, 41–50 (2015).

27. Okada, M. & Shi, Y. B. The balance of two opposing factors Mad and Myc regulates cell fate during tissue remodeling. *Cell Biosci.* **8**, 1–9 (2018).

28. Lüscher, B. MAD1 and its life as a MYC antagonist: An update. *European Journal of Cell Biology* **91**, 506–514 (2012).

29. Meran, L., Baulies, A. & Li, V. S. W. Intestinal Stem Cell Niche: The Extracellular Matrix and Cellular Components. *Stem Cells Int.* **2017**, 1–11 (2017).

30. Batlle, E. *et al.* β-Catenin and TCF Mediate Cell Positioning in the Intestinal Epithelium by Controlling the Expression of EphB/EphrinB. *Cell* **111**, 251–263 (2002).

31. Vogelstein, B. & Kinzler, K. W. The multistep nature of cancer. *Trends Genet.* **9**, 138–141 (1993).

32. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).

33. Asfaha, S. *et al.* Krt19+/Lgr5− Cells Are Radioresistant Cancer-Initiating Stem Cells in the Colon and Intestine. *Cell Stem Cell* **16**, 627–638 (2015).

34. Krause, M., Dubrovska, A., Linge, A. & Baumann, M. Cancer stem cells: Radioresistance, prediction of radiotherapy outcome and specific targets for combined treatments. *Adv. Drug Deliv. Rev.* **109**, 63–73 (2017).

35. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).

36. Sangiorgi, E. & Capecchi, M. R. Bmi1 is expressed in vivo in intestinal stem cells. *Nat. Genet.* **40**, 915–920 (2008).

37. Hatano, Y. *et al.* Multifaceted interpretation of colon cancer stem cells. *Int. J. Mol. Sci.* **18**, (2017).

38. Tian, H. *et al.* A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478**, 255–259 (2011).

39. Cole, J. W. & McKalen, A. Studies on the morphogenesis of adenomatous polyps in the human colon. *Cancer* **16**, 998–1002 (1963).

40. Shih, I. *et al.* Top-down morphogenesis of colorectal tumors. *PNAS* **98**, 2640–2645 (2001).

41. Schwitalla, S. *et al.* Intestinal Tumorigenesis Initiated by Dedifferentiation and Acquisition of Stem-Cell-like Properties. *Cell* **152**, 25–38 (2013).

42. Wong, W.-M. *et al.* Histogenesis of human colorectal adenomas and hyperplastic polyps: the role of cell proliferation and crypt fission. *Gut* **50**, 212–217 (2002).

43. Benedix, F. *et al.* Comparison of 17,641 patients with right- and left-sided colon cancer: Differences in epidemiology, perioperative course, histology, and survival. *Dis. Colon Rectum* **53**, 57–64 (2010).

44. Lee, G. H. *et al.* Is right-sided colon cancer different to left-sided colorectal cancer? – A systematic review. *Eur. J. Surg. Oncol.* **41**, 300–308 (2015).

45. Gonzalez, E. C., Roetzheim, R. G., Ferrante, J. M. & Campbell, R. Predictors of proximal vs. distal colorectal cancers. *Dis. Colon Rectum* **44**, 251–258 (2001).

46. Meguid, R. A., Slidell, M. B., Wolfgang, C. L., Chang, D. C. & Ahuja, N. Is there a difference in survival between right- versus left-sided colon cancers? *Ann. Surg. Oncol.* **15**, 2388–2394 (2008).

47. Ooi, L. Y. Post-GWAS functional characterisation of colorectal cancer risk loci. (Edinburgh, 2016).

48. Nawa, T. *et al.* Differences between right- and left-sided colon cancer in patient characteristics, cancer morphology and histology. *J. Gastroenterol. Hepatol.* **23**, 418–423 (2008).

49. Kim, S.-E. *et al.* Sex- and gender-specific disparities in colorectal cancer risk. *World J. Gastroenterol.* **21**, 5167–75 (2015).

50. Kang, H. *et al.* Rare tumors of the colon and rectum: a national review. *Int. J. Colorectal Dis.* **22**, 183–189 (2006).

51. Poulogiannis, G. *et al.* Prognostic relevance of DNA copy number changes in colorectal cancer. *J. Pathol.* **220**, 338–347 (2010).

52. Pino, M. S. & Chung, D. C. The Chromosomal Instability Pathway in Colon Cancer. *Gastroenterology* **138**, 2059–2072 (2010).

53. Bowtell, D. D. *et al.* Rethinking ovarian cancer II: Reducing mortality from high-grade serous ovarian cancer. *Nature Reviews Cancer* (2015). doi:10.1038/nrc4019

54. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).

55. Müller, M. F., Ibrahim, A. E. K. & Arends, M. J. Molecular pathological classification of colorectal cancer. *Virchows Arch. Eur. J. Pathol.* **469**, 125–134 (2016).

56. Vogelstein, B. *et al.* Genetic Alterations during Colorectal-Tumor Development. *N. Engl. J. Med.* **319**, 525–532 (1988).

57. Silva, A.-L. *et al.* Boosting Wnt activity during colorectal cancer progression through selective hypermethylation of Wnt signaling antagonists. *BMC Cancer* **14**, 891 (2014).

58. Jianjiong, G. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

59. Ibrahim, A. E. K. *et al.* Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. *Gut* **60**, 499–508 (2011).

60. Boland, C. R. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* **138**, 2073-2087.e3 (2010).

61. Rustgi, A. K. The genetics of hereditary colon cancer. *Genes Dev.* **21**, 2525–38 (2007).

62. Liu, D., Keijzers, G. & Rasmussen, L. J. DNA mismatch repair and its many roles in eukaryotic cells. *Mutat. Res. - Rev. Mutat. Res.* **773**, 174–187 (2017).

63. Albertson, T. M. *et al.* DNA polymerase and proofreading suppress discrete mutator and cancer phenotypes in mice. *Proc. Natl. Acad. Sci.* **106**, 17101–17104 (2009).

64. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–4 (2012).

65. Cantwell-Dorris, E. R., O'Leary, J. J. & Sheils, O. M. BRAFV600E: implications for carcinogenesis and molecular therapy. *Mol. Cancer Ther.* **10**, 385–94 (2011).

66. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

67. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).

68. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).

69. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

70. Llosa, N. J. *et al.* The Vigorous Immune Microenvironment of Microsatellite Instable Colon Cancer Is Balanced by Multiple Counter-Inhibitory Checkpoints. *Cancer Discov.* **5**, 43–51 (2015).

71. Walther, A., Johnstone, E., Swanton, C. & Midgley, R. Genetic prognostic and predictive markers in colorectal cancer. (2011). doi:10.1038/nrc2645

72. Gelsomino, F., Barbolini, M., Spallanzani, A., Pugliese, G. & Cascinu, S. The evolving role of microsatellite instability in colorectal cancer: A review. *Cancer Treat. Rev.* **51**, 19–26 (2016).

73. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

74. Colle, R. *et al.* Immunotherapy and patients treated for cancer with microsatellite instability. *Bull. Cancer* **104**, 42–51 (2017).

75. McConechy, M. K. *et al.* Endometrial Carcinomas with POLE Exonuclease Domain Mutations Have a Favorable Prognosis. *Clin. Cancer Res.* **22**, 2865–2873 (2016).

76. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17**, 79–92 (2017).

77. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum. Mol. Genet.* **23**, 4729–4737 (2014).

78. Lynch, H. T. & Lynch, J. F. Genetics of colonic cancer. *Digestion* **59**, 481–492 (1998).

79. Wennstrom, J., Pierce, E. R. & McKusick, V. A. Hereditary benign and malignant lesions of the large bowel. *Cancer* **34**, 850–857 (1974).

80. Bisgaard, M. L., Fenger, K., Bülow, S., Niebuhr, E. & Mohr, J. Familial adenomatous polyposis (FAP): Frequency, penetrance, and mutation rate. *Hum. Mutat.* **3**, 121–125 (1994).

81. Aihara, H., Kumar, N. & Thompson, C. C. Diagnosis, surveillance, and treatment strategies for familial adenomatous polyposis: rationale and update. *Eur. J. Gastroenterol. Hepatol.* **26**, 255–62 (2014).

82. Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589–600 (1991).

83. Krishnamurthy, N. & Kurzrock, R. Targeting the Wnt/beta-catenin pathway in cancer: Update on effectors and inhibitors. *Cancer Treat. Rev.* **62**, 50–60 (2018).

84. Tai, D. *et al.* Targeting the WNT Signaling Pathway in Cancer Therapeutics. *Oncologist* **20**, 1189–1198 (2015).

85. Knudson, A. G. Hereditary cancer: Two hits revisited. *J. Cancer Res. Clin. Oncol.* **122**, 135–140 (1996).

86. Rowan, A. J. *et al.* APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3352–7 (2000).

87. Lynch, H. T. & de la Chapelle, A. Hereditary colorectal cancer. *N.Engl.J Med* **348**, 919–932 (2003).

88. Mitchell, R. J., Farrington, S. M., Dunlop, M. G. & Campbell, H. Mismatch Repair Genes hMLH1 and hMSH2 and Colorectal Cancer: A HuGE Review. *Am. J. Epidemiol.* **156**, 885–902 (2002).

89. Dunlop, M. G. *et al.* Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum. Mol. Genet.* **6**, 105–110 (1997).

90. Lin, K. M. *et al.* Cumulative incidence of colorectal and extracolonic cancers in MLH1 and MSH2 mutation carriers of hereditary nonpolyposis colorectal cancer. *J. Gastrointest. Surg.* **2**, 67–71 (1998).

91. Watson, P. & Lynch, H. T. Extracolonic cancer in hereditary nonpolyposis colorectal cancer. *Cancer* **71**, 677–685 (1993).

92. Roncucci, L., Pedroni, M. & Mariani, F. Attenuated adenomatous polyposis of the large bowel: Present and future. *World J. Gastroenterol.* **23**, 4135–4139 (2017).

93. Ngeow, J. *et al.* Prevalence of germline PTEN, BMPR1A, SMAD4, STK11, and ENG mutations in patients with moderate-load colorectal polyps. *Gastroenterology* **144**, 1402-1409.e5 (2013).

94. Eng, C. Will the real Cowden syndrome please stand up: revised diagnostic criteria. *J.*

Med. Genet. **37**, 828–30 (2000).

95. Brosens, L. A. A. *et al.* Risk of colorectal cancer in juvenile polyposis. *Gut* **56**, 965–7 (2007).

96. Law, P. J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat. Commun.* **10**, 1–15 (2019).

97. Pasche, B. *et al.* TGFBR1*6A and cancer: a meta-analysis of 12 case-control studies. *J. Clin. Oncol.* **22**, 756–8 (2004).

98. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk [Letter]. *Nat. Genet.* **44**, 770–776 (2012).

99. Botteri, E. *et al.* Smoking and Colorectal Cancer. *Jama* **300**, 2765 (2008).

100. Jayasekara, H., MacInnis, R. J., Room, R. & English, D. R. Long-Term Alcohol Consumption and Breast, Upper Aero-Digestive Tract and Colorectal Cancer Risk: A Systematic Review and Meta-Analysis. *Alcohol Alcohol.* **51**, 315–330 (2016).

101. Bastide, N. M., Pierre, F. H. F. & Corpet, D. E. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev. Res. (Phila).* **4**, 177–84 (2011).

102. Alexander, D. D., Weed, D. L., Miller, P. E. & Mohamed, M. A. Red Meat and Colorectal Cancer: A Quantitative Update on the State of the Epidemiologic Science. *J. Am. Coll. Nutr.* **34**, 521–543 (2015).

103. McIntyre, A., Gibson, P. R. & Young, G. P. Butyrate production from dietary fibre and protection against large bowel cancer in a rat model. *Gut* **34**, 386–91 (1993).

104. Zeng, H., Lazarova, D. L. & Bordonaro, M. Mechanisms linking dietary fiber, gut microbiota and colon cancer prevention. *World J. Gastrointest. Oncol.* **6**, 41 (2014).

105. Jayasekara, H. *et al.* Associations of alcohol intake, smoking, physical activity and obesity with survival following colorectal cancer diagnosis by stage, anatomic site and tumor molecular subtype. *Int. J. Cancer* **142**, 238–250 (2018).

106. Sanchez, N. F. *et al.* Physical activity reduces risk for colon polyps in a multiethnic colorectal cancer screening population. *BMC Res. Notes* **5**, 312 (2012).

107. Schöttker, B. *et al.* Strong associations of 25-hydroxyvitamin D concentrations with all-cause, cardiovascular, cancer, and respiratory disease mortality in a large cohort study. *Am. J. Clin. Nutr.* **97**, 782–793 (2013).

108. Zgaga, L. *et al.* Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer. *J. Clin. Oncol.* **32**, 2430–2439 (2014).

109. He, Y. *et al.* Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: A large Mendelian randomisation study. *BMC Med.* **16**, 1–11 (2018).

110. Procopciuc, L. M., Osian, G. & Iancu, M. N-acetyl transferase 2/environmental factors and their association as a modulating risk factor for sporadic colon and rectal cancer. *J. Clin. Lab. Anal.* **31**, e22098 (2017).

111. Ritz, B. R. *et al.* Lessons Learned From Past Gene-Environment Interaction Successes. *Am. J. Epidemiol.* **186**, 778–786 (2017).

112. Ricciotti, E. & FitzGerald, G. A. Prostaglandins and inflammation. *Arterioscler. Thromb. Vasc. Biol.* **31**, 986–1000 (2011).

113. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, Inflammation, and Cancer. *Cell* **140**, 883–899 (2010).

114. Munkholm, P. Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **18**, 1–5 (2003).

115. von Roon, A. C. *et al.* The Risk of Cancer in Patients with Crohn's Disease. *Dis. Colon Rectum* **50**, 839–855 (2007).

116. Rothwell, P. M. *et al.* Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *Lancet* **377**, 31–41 (2011).

117. Tang, J., Sharif, O., Pai, C. & Silverman, A. L. Mesalamine protects against colorectal cancer in inflammatory bowel disease. *Dig. Dis. Sci.* **55**, 1696–1703 (2010).

118. Din, F. V. N. *et al.* Aspirin inhibits mTOR signaling, activates AMP-activated protein kinase, and induces autophagy in colorectal cancer cells. *Gastroenterology* **142**, 1504–1515 (2012).

119. Sun, D. *et al.* Aspirin disrupts the mTOR-Raptor complex and potentiates the anti-cancer activities of sorafenib via mTORC1 inhibition. *Cancer Lett.* **406**, 105–115 (2017).

120. Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* **66**, 633–643 (2017).

121. Engin, A. B., Karahalil, B., Karakaya, A. E. & Engin, A. *Helicobacter pylori* and serum kynurenine-tryptophan ratio in patients with colorectal cancer. *World J. Gastroenterol.* **21**, 3636 (2015).

122. Chen, J., Pitmon, E. & Wang, K. Microbiome, inflammation and colorectal cancer. *Semin. Immunol.* **32**, 43–53 (2017).

123. Mima, K. *et al.* Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut* **65**, 1973–1980 (2016).

124. Ye, X. *et al. Fusobacterium Nucleatum* Subspecies *Animalis* Influences Proinflammatory Cytokine Expression and Monocyte Activation in Human Colorectal Tumors. *Cancer Prev. Res.* **10**, 398–409 (2017).

125. Purcell, R. V, Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.* **7**, 11590 (2017).

126. Smith, P. M. *et al.* The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. *Science (80-.).* **341**, 569–573 (2013).

127. Chang, P. V, Hao, L., Offermanns, S. & Medzhitov, R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2247–52 (2014).

128. Lichtenstein, P. Environmental and heritable factors in the causation of cancer -

Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).

129. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* **99**, 260–266 (2002).

130. Jiao, S. *et al.* Estimating the heritability of colorectal cancer. *Hum. Mol. Genet.* **23**, 3898–3905 (2014).

131. Muñoz, M. *et al.* Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nat. Genet.* **48**, 980–983 (2016).

132. Corradin, O. *et al.* Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat. Genet.* **48**, (2016).

133. Romiguier, J., Ranwez, V., Douzery, E. J. P. & Galtier, N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* **20**, 1001–1009 (2010).

134. Mercer, T. R. *et al.* Genome-wide discovery of human splicing branchpoints. *Genome Res.* **25**, 290–303 (2015).

135. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).

136. Editors. Intron sequence. *Biol. Dict.* (2017).

137. Kim, T. K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).

138. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).

139. Williamson, I., Lettice, L. A., Hill, R. E. & Bickmore, W. A. Shh and ZRS enhancer

colocalisation is specific to the zone of polarising activity. *Development* **143**, 2994–3001 (2016).

140. Aerts, S. Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. *Curr. Top. Dev. Biol.* **98**, 121–145 (2012).

141. Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. & Kadonaga, J. T. The RNA polymerase II core promoter — the gateway to transcription. *Curr. Opin. Cell Biol.* **20**, 253–259 (2008).

142. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111**, 6131–6138 (2014).

143. Hnisz, D. *et al.* Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Mol. Cell* **58**, 362–370 (2015).

144. Hnisz, D. *et al.* XSuper-enhancers in the control of cell identity and disease. *Cell* **155**, 934 (2013).

145. Schultz, M. C., Reeder, R. H. & Hahn, S. Variants of the TATA-binding protein can distinguish subsets of RNA polymerase I, II, and III promoters. *Cell* **69**, 697–702 (1992).

146. Sikorski, T. W. & Buratowski, S. The basal initiation machinery: beyond the general transcription factors. *Curr. Opin. Cell Biol.* **21**, 344–351 (2009).

147. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).

148. Liu, X., Bushnell, D. A. & Kornberg, R. D. RNA polymerase II transcription: Structure and mechanism. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 2–8 (2013).

149. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).

150. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).

151. Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* **35**, 125–131 (2007).

152. Herzel, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.* **18**, 637–650 (2017).

153. Fuller-pace, F. V. DExD / H box RNA helicases : multifunctional proteins with important roles in transcriptional regulation. **34**, 4206–4215 (2006).

154. Cretu, C. *et al.* Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Mol. Cell* **64**, 307–319 (2016).

155. Butcher, S. E. The spliceosome as ribozyme hypothesis takes a second step. *Proc. Natl. Acad. Sci.* **106**, 12211–12212 (2009).

156. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).

157. Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**, e268 (2004).

158. Berglund, J. A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. The Splicing Factor BBP Interacts Specifically with the Pre-mRNA Branchpoint Sequence UACUAAC. *Cell* **89**, 781–787 (1997).

159. Shao, C. *et al.* Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. *Nat. Struct. Mol. Biol.* **21**, 997–1005 (2014).

160. Perea, W., Schroeder, K. T., Bryant, A. N. & Greenbaum, N. L. Interaction between the Spliceosomal Pre-mRNA Branch Site and U2 snRNP Protein p14. *Biochemistry* **55**, 629–632 (2016).

161. Machyna, M., Heyn, P. & Neugebauer, K. M. Cajal bodies: where form meets function. *Wiley Interdiscip. Rev. RNA* **4**, 17–34 (2013).

162. Bertram, K. *et al.* Cryo-EM structure of a human spliceosome activated for step 2 of

splicing. *Nature* **542**, 318–323 (2017).

163. Hegele, A. *et al.* Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome. *Mol. Cell* **45**, 567–580 (2012).

164. Small, E. C., Leggett, S. R., Winans, A. A. & Staley, J. P. The EF-G-like GTPase Snu114p Regulates Spliceosome Dynamics Mediated by Brr2p, a DExD/H Box ATPase. *Mol. Cell* **23**, 389–399 (2006).

165. Boesler, C. *et al.* Stable tri-snRNP integration is accompanied by a major structural rearrangement of the spliceosome that is dependent on Prp8 interaction with the 5' splice site. *RNA* **21**, 1993–2005 (2015).

166. Tsai, R.-T. *et al.* Dynamic interactions of Ntr1-Ntr2 with Prp43 and with U5 govern the recruitment of Prp43 to mediate spliceosome disassembly. *Mol. Cell. Biol.* **27**, 8027–37 (2007).

167. Zheng, Z.-M. Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J. Biomed. Sci.* **11**, 278–94 (2004).

168. Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C. & Black, D. L. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.* **15**, 183–191 (2008).

169. Izquierdo, J. M. *et al.* Regulation of fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell* **19**, 475–484 (2005).

170. Attanasio, C., David, A. & Neerman-Arbez, M. Outcome of donor splice site mutations accounting for congenital afibrinogenemia reflects order of intron removal in the fibrinogen alpha gene (FGA). *Blood* **101**, 1851–6 (2003).

171. Pandit, S. *et al.* Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. *Mol. Cell* **50**, 223–235 (2013).

172. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–

1052 (2012).

173. Hastings, M. L., Allemand, E., Duelli, D. M., Myers, M. P. & Krainer, A. R. Control of Pre-mRNA Splicing by the General Splicing Factors PUF60 and U2AF65. *PLoS One* **2**, e538 (2007).

174. Soemedi, R. *et al.* The effects of structure on pre-mRNA processing and stability. *Methods* **125**, 36–44 (2017).

175. Zhao, C. & Pyle, A. M. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem. Sci.* **42**, 470–482 (2017).

176. Auboeuf, D. *et al.* Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2270–4 (2004).

177. de la Mata, M., Lafaille, C. & Kornblihtt, A. R. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA* **16**, 904–12 (2010).

178. de la Mata, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* **12**, 525–532 (2003).

179. Dujardin, G. *et al.* How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Mol. Cell* **54**, 683–690 (2014).

180. Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* **16**, 996–1001 (2009).

181. Huff, J. T., Zilberman, D. & Roy, S. W. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538**, 533–536 (2016).

182. Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal Fluctuations Govern the Transcription Dynamics of RNA Polymerase II. *Science (80-.).* **325**, 626–628 (2009).

183. Sims, R. J. *et al.* Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. *Mol. Cell*

28, 665–676 (2007).

184. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).

185. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).

186. Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R. & Bickmore, W. A. Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.* **8**, (2012).

187. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144**, 16–26 (2011).

188. Aregger, M. & Cowling, V. H. Regulation of mRNA capping in the cell cycle. *RNA Biol.* **14**, 11–14 (2017).

189. Ramanathan, A., Robb, G. B. & Chan, S. H. mRNA capping: Biological functions and applications. *Nucleic Acids Res.* **44**, 7511–7526 (2016).

190. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA* **3**, 385–396 (2012).

191. Millevoi, S. & Vagner, S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res.* **38**, 2757–2774 (2009).

192. Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11–23 (2001).

193. Shi, Y. & Manley, J. L. The end of the message: Multiple protein–RNA interactions define the mRNA polyadenylation site. *Genes Dev.* **29**, 889–897 (2015).

194. Engel, C., Sainsbury, S., Cheung, A. C., Kostrewa, D. & Cramer, P. RNA polymerase I structure and transcription regulation. *Nature* **502**, 650–655 (2013).

195. Han, Y., Yan, C., Fishbain, S., Ivanov, I. & He, Y. Cell Discovery Structural visualization of RNA polymerase III transcription machineries. *Cell Discov.* **4**, 40 (2018).

196. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science (80-.).* **352**, 1413–1416 (2016).

197. Kobayashi, H. & Tomari, Y. RISC assembly: Coordination between small RNAs and Argonaute proteins. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1859**, 71–81 (2016).

198. Baltz, A. G. *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol. Cell* **46**, 674–690 (2012).

199. Taliaferro, J. M. *et al.* RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Mol. Cell* **64**, 294–306 (2016).

200. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).

201. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6 (2008).

202. Xu, S. *et al.* Editing aberrant splice sites efficiently restores β-globin expression in β-thalassemia. *Blood* **133**, 2255–2262 (2019).

203. Takeshima, Y. *et al.* Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. *J. Hum. Genet.* **55**, 379–388 (2010).

204. Samaranch, L. *et al.* PINK1-linked parkinsonism is associated with Lewy body pathology. *Brain* **133**, 1128–1142 (2010).

205. OTOMO, J. *et al.* Electrophysiological and Histopathological Characteristics of Progressive Atrioventricular Block Accompanied by Familial Dilated Cardiomyopathy Caused by a Novel Mutation of Lamin A/C Gene. *J. Cardiovasc. Electrophysiol.* **16**, 137–145 (2005).

206. Morel, C. F. *et al.* A *LMNA* Splicing Mutation in Two Sisters with Severe Dunnigan-Type Familial Partial Lipodystrophy Type 2. *J. Clin. Endocrinol. Metab.* **91**, 2689–2695 (2006).

207. Muchir, A. *et al.* Identification of mutations in the gene encoding lamins A/C in

autosomal dominant limb girdle muscular dystrophy with atrioventricular conduction disturbances (LGMD1B). *Hum. Mol. Genet.* **9**, 1453–1459 (2000).

208. De Sandre-Giovannoli, A. *et al.* Lamin a truncation in Hutchinson-Gilford progeria. *Science* **300**, 2055 (2003).

209. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

210. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

211. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

212. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *bioRxiv* **32**, 020255 (2015).

213. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science (80-.).* **296**, 2225–9 (2002).

214. Kindt, A. S. D., Navarro, P., Semple, C. A. M. & Haley, C. S. The genomic signature of trait-associated variants. *BMC Genomics* **14**, 1 (2013).

215. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

216. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 1–14 (2019). doi:10.1016/j.cell.2018.12.015

217. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).

218. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 45–63 (2017).

219. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**,

1747–1751 (2017).

220. Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genet.* **11**, 1–29 (2015).

221. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).

222. Singh, T. *et al.* Characterization of expression quantitative trait loci in the human colon. *Inflamm. Bowel Dis.* **21**, 251–6 (2015).

223. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

224. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).

225. Milne, R. L. & Antoniou, A. C. Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Ann. Oncol.* **22**, 11–17 (2011).

226. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–35 (2008).

227. Al-Tassan, N. A. *et al.* A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci. Rep.* **5**, 10442 (2015).

228. Tenesa, A. & Dunlop, M. G. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat. Rev. Genet.* **10**, 353–358 (2009).

229. Timofeeva, M. N. *et al.* Recurrent Coding Sequence Variation Explains Only A Small Fraction of the Genetic Architecture of Colorectal Cancer. *Sci. Rep.* **5**, 16286 (2015).

230. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).

231. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2018).

232. Jiang, H. *et al.* Long non-coding RNA SNHG15 interacts with and stabilizes transcription factor Slug and promotes colon cancer progression. *Cancer Lett.* **425**, 78–87 (2018).

233. Shan, Z. *et al.* Long non-coding RNA Linc00675 suppresses cell proliferation and metastasis in colorectal cancer via acting on miR-942 and Wnt/β-catenin signaling. *Biomed. Pharmacother.* **101**, 769–776 (2018).

234. Sun, Y., Zheng, Z.-P., Li, H., Zhang, H.-Q. & Ma, F.-Q. ANRIL is associated with the survival rate of patients with colorectal cancer, and affects cell migration and invasion in vitro. *Mol. Med. Rep.* **14**, 1714–1720 (2016).

235. Sun, Z. *et al.* Downregulation of long non-coding RNA ANRIL suppresses lymphangiogenesis and lymphatic metastasis in colorectal cancer. *Oncotarget* **7**, 47536–47555 (2016).

236. Orlando, G. *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum. Mol. Genet.* **25**, 2349–2359 (2016).

237. Zhang, L. *et al.* Systematic identification of cancer-related long noncoding RNAs and aberrant alternative splicing of quintuple-negative lung adenocarcinoma through RNA-Seq. *Lung Cancer* **109**, 21–27 (2017).

238. Zhao, D. *et al.* Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-deficient cancer. *Nature* **542**, 484–488 (2017).

239. Kostic, A. D., Xavier, R. J. & Gevers, D. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. *Gastroenterology* **146**, 1489–1499 (2014).

240. Hall, A. B., Tolonen, A. C. & Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* **18**, 690–699 (2017).

241. Zeng, Z. B., Robertson, A., Hewitt, J. D., Zamir, D. & Rabinowitch, H. D. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–68 (1994).

242. Hewitson, P., Glasziou, P. P., Irwig, L., Towler, B. & Watson, E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. *Cochrane Database Syst. Rev.* (2007). doi:10.1002/14651858.CD001216.pub2

243. Vasen, H. F. A. *et al.* One to 2-Year Surveillance Intervals Reduce Risk of Colorectal Cancer in Families With Lynch Syndrome. *Gastroenterology* **138**, 2300–2306 (2010).

244. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).

245. He, Y. *et al.* Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. *Int. J. Cancer* **145**, 2427–2432 (2019).

246. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research* **5**, 1356 (2016).

247. Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* **5**, 4698 (2014).

248. LEHMANN, K.-V. *et al.* Integrative genome-wide analysis of the determinants of rna splicing in kidney renal clear cell carcinoma. in *Biocomputing 2015* 44–55 (2015). doi:10.1142/9789814644730_0006

249. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).

250. Pickrell, J., Marioni, J., Pai, A., Degner, J. F. & Engelhardt, B. E. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

251. Zhao, K., Lu, Z., Park, J. W., Zhou, Q. & Xing, Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.* **14**, R74 (2013).

252. Jia, C., Hu, Y., Liu, Y. & Li, M. Mapping splicing quantitative trait loci in RNA-seq. *Cancer Inform.* **14**, 45–53 (2014).

253. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).

254. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).

255. Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Stat.* **5**, 445–463 (1977).

256. Anderson, M. & J. A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46 (2001).

257. Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* **3**, e01381 (2014).

258. Ongen, H. & Dermitzakis, E. T. Alternative Splicing QTLs in European and African Populations. *Am. J. Hum. Genet.* **97**, 567–575 (2015).

259. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 1–7 (2015). doi:10.1093/bioinformatics/btv722

260. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).

261. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science (80-.).* **352**, 600–04 (2016).

262. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

263. Scelo, G. *et al.* Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* **8**, 15724 (2017).

264. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8**, 14519 (2017).

265. Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0238-1

266. Malik, M. *et al.* CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. *J. Neurosci.* **33**, 13320–5 (2013).

267. Raj, T. *et al.* CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. *Hum. Mol. Genet.* **23**, 2729–2736 (2014).

268. Nixon, R. A. New perspectives on lysosomes in ageing and neurodegenerative disease. *Ageing Research Reviews* **32**, 1 (2016).

269. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).

270. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

271. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–58 (2013).

272. Seiler, M. *et al.* Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep.* **23**, 282-296.e4 (2018).

273. Kim, E. *et al.* SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27**, 617–630 (2015).

274. Zhao, J. *et al.* Functional analysis reveals that RBM10 mutations contribute to lung adenocarcinoma pathogenesis by deregulating splicing. *Sci. Rep.* **7**, 40488 (2017).

275. Hernández, J. *et al.* Tumor suppressor properties of the splicing regulatory factor RBM10. *RNA Biol.* **13**, 466–472 (2016).

276. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell* **156**, 1324–1335 (2014).

277. Mahalanobis, P. C. On the Generalised Distance in Statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936).

278. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

279. Cuajungco, M. P. *et al.* Abnormal accumulation of human transmembrane (TMEM)-176A and 176B proteins is associated with cancer pathology. *Acta Histochem.* **114**, 705–712 (2012).

280. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.* **20**, 2215–2226 (2017).

281. Zhou, C. *et al.* The Rac1 splice form Rac1b promotes K-ras-induced lung tumorigenesis. *Oncogene* **32**, 903–909 (2013).

282. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).

283. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* **7**, 1–13 (2015).

284. Shiraishi, Y. *et al.* A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Doi.Org* 162560 (2017). doi:10.1101/162560

285. Nishida, A. *et al.* Chemical treatment enhances skipping of a mutated exon in the dystrophin gene. *Nat. Commun.* **2**, 308 (2011).

286. Ma, P. C. *et al.* Functional expression and mutations of c-Met and its therapeutic inhibition with SU11274 and small interfering RNA in non-small cell lung cancer. *Cancer Res.* **65**, 1479–1488 (2005).

287. Schödel, J. *et al.* Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat. Genet.* **44**, 420–425 (2012).

288. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418–421 (2015).

289. Stacey, S. N. *et al.* New basal cell carcinoma susceptibility loci. *Nat. Commun.* **6**, 6825 (2015).

290. Mesrian Tanha, H., Rahgozar, S. & Mojtabavi Naeini, M. ABCC4 functional SNP in the 3' splice acceptor site of exon 8 (G912T) is associated with unfavorable clinical outcome in children with acute lymphoblastic leukemia. *Cancer Chemother. Pharmacol.* **80**, 109–117 (2017).

291. Soukarieh, O. *et al.* Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. *PLOS Genet.* **12**, e1005756 (2016).

292. Rhine, C. L. *et al.* Hereditary cancer genes are highly susceptible to splicing mutations. *PLOS Genet.* **14**, e1007231 (2018).

293. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).

294. Huang, H. H. *et al.* Proteasome inhibitor-induced modulation reveals the spliceosome as a specific therapeutic vulnerability in multiple myeloma. *bioRxiv* (2018). doi:10.1101/508549

295. Hsu, T. Y.-T. *et al.* The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nat. Lett.* **525**, 384–388 (2015).

296. Van Alphen, R. J., Wiemer, E. A. C., Burger, H. & Eskens, F. A. L. M. The spliceosome as target for anticancer treatment. *British Journal of Cancer* **100**, 228–232 (2009).

297. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211-224.e6 (2018).

298. Zgaga, L. *et al.* Diet, Environmental Factors, and Lifestyle Underlie the High Prevalence of Vitamin D Deficiency in Healthy Adults in Scotland, and Supplementation Reduces the Proportion That Are Severely Deficient. *J. Nutr.* **141**, 1535–1542 (2011).

299. Theodoratou, E. *et al.* Instrumental Variable Estimation of the Causal Effect of Plasma 25-Hydroxy-Vitamin D on Colorectal Cancer Risk: A Mendelian Randomization Analysis. *PLoS One* **7**, e37662 (2012).

300. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

301. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *Nat. Methods* **14**, 417–419 (2017).

302. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

303. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–5 (2010).

304. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

305. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

306. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

307. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131–e131 (2010).

308. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

309. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

310. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: A rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32**, i192–i200 (2016).

311. Bohnert, R. & Ratsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38**, W348–W351 (2010).

312. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

313. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).

314. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).

315. O'Neil, D., Glowatz, H. & Schlumpberge, M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.* **103**, 4.19.1-4.19.8 (2013).

316. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 1–11 (2014).

317. Kher, G., Trehan, S. & Ambikanandan, M. Antisense Oligonucleotides and RNA Interference. in *Challenges in Delivery of Therapeutic Genomics and Proteomics* (ed. Misra, A.) 325–386 (Elsevier, 2011).

318. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS One* **7**, e42882 (2012).

319. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**, 83–90 (2017).

320. Zheng-Bradley, X. *et al.* Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* **6**, (2017).

321. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 1–43 (2005).

322. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

323. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).

324. Love, M. I., Hogenesch, J. B. & Irizarry, R. A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* **34**, 1287–1291 (2016).

325. Broad Institute. Picard Toolkit version 1.139. (2015).

326. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

327. Ruffier, M. *et al.* Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *Database* **2017**, (2017).

328. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).

329. Phillippy, A. M. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

330. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. in *Current Protocols in Bioinformatics* **43**, 11.10.1-11.10.33 (John Wiley & Sons, Inc., 2013).

331. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

332. Stults, D. M., Killen, M. W., Pierce, H. H. & Pierce, A. J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **18**, 13–18 (2008).

333. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).

334. Jones, D. cgpBigWig. (2017).

335. Bolstad, B. M., Irizarry, R. ., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

336. Mardia, K. V., Kent, J. T. & Bibby, J. M. *Multivariate Analysis*. (London: Academic Press, 1979).

337. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer-Verlag, 2002).

338. Becker, R. A., Chambers, J. M. & Wilks, A. R. *The New S Language*. (Wadsworth & Brooks/Cole, 1988).

339. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

340. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

341. Leek, J. T. *et al.* sva: Surrogate Variable Analysis. (2015).

342. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

343. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, (2010).

344. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

345. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

346. Aulchenko, Y. S., Struchalin, M. V & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).

347. Stegle, O., Parts, L., Durbin, R. & Winn, J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, 1–11 (2010).

348. Parts, L., Stegle, O., Winn, J. & Durbin, R. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet.* **7**, 1–10 (2011).

349. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

350. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Rev.* **51**, 661–703 (2009).

351. De Las Rivas, J. & Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput. Biol.* **6**, e1000807 (2010).

352. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

353. Müllner, D. Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *J. Stat. Softw.* **53**, 1–18 (2013).

354. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: An introduction to cluster analysis*. (John Wiley Sons, Inc, 1990).

355. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

356. Zhang, C., Zhang, B., Lin, L. L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 1–11 (2017).

357. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data

with or without a reference genome. *BMC Bioinformatics* (2011).

358. Smith, T. Why you should use alignment-independent quantification for RNA-Seq. *CGAT Oxford Blog* (2016).

359. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).

360. Mondal, A. M. *et al.* p53 isoforms regulate aging- and tumor-associated replicative senescence in T lymphocytes. *J. Clin. Invest.* **123**, 5247–5257 (2013).

361. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).

362. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15**, 1–15 (2014).

363. Hembach, K. *et al.* RNA sequencing data : hitchhiker ' s guide to expression analysis. *PeerJ Prepr.* (2018). doi:10.7287/peerj.preprints.27283v1

364. Monlong, J. sQTLseekeR development page. *GitHub* (2018).

365. Fatima, A. *et al.* Weighted Gene Co-Expression Network Analysis Identifies Gender Specific Modules and Hub Genes Related to Metabolism and Inflammation in Response to an Acute Lipid Challenge. *Mol. Nutr. Food Res.* **62**, 1–8 (2018).

366. Campbell, C. L. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

367. Kuhn, R. M., Haussler, D. & James Kent, W. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

368. Storey, J. D. False Discovery Rates. *Princet. Univ. Princeton, USA* 1–7 (2010). doi:10.1198/016214507000000941

369. Jones, E., Oliphant, T. & Pearu, P. SciPy: Open source scientific tools for Python. (2001).

370. Gilad, Y. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

371. Bonferroni, C. E. Statistical Theory of Classes and Probability. *R Publ. Inst. Econ. Commer. Sci. Florence Florence, Italy* **8**, 3–62 (1936).

372. Hochberg, Y. & Benjamini, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

373. Norman, P. J. *et al.* Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class i and II. *Genome Res.* **27**, 813–823 (2017).

374. Kennedy, A. E., Ozbek, U. & Dorak, M. T. What has GWAS done for HLA and disease associations? *Int. J. Immunogenet.* **44**, 195–211 (2017).

375. Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* **4**, (2008).

376. Foissac, S. & Sammeth, M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* **35**, W297-9 (2007).

377. Turner, S. & D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* (2014). doi:10.1101/005165

378. Van De Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

379. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15**, 467 (2014).

380. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).

381. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512

(2013).

382. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

383. Stein, S., Lu, Z., Bahrami-Samani, E., Park, J. W. & Xing, Y. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Res.* **43**, 10612–10622 (2015).

384. Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol.* **16**, 131 (2015).

385. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).

386. Stranger, B. E. *et al.* Population genomics of human gene expression. **39**, 1217–1224 (2007).

387. Kahles, A., Ong, C. S., Zhong, Y. & Rätsch, G. SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**, 1840–1847 (2016).

388. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).

389. Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* **27**, 1759–1768 (2017).

390. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178

391. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).

392. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

393. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* **9**, 490 (2018).

394. Musella, V. *et al.* Effects of Warm Ischemic Time on Gene Expression Profiling in Colorectal Cancer Tissues and Normal Mucosa. *PLoS One* **8**, e53406 (2013).

395. Qu, W., Gurdziel, K., Pique-Regi, R. & Ruden, D. M. Identification of splicing quantitative trait loci (sQTL) in Drosophila melanogaster with developmental lead (Pb2+) exposure. *Front. Genet.* **8**, 1–12 (2017).

396. Forrest, M. E. *et al.* Colon Cancer-Upregulated Long Non-Coding RNA lincDUSP Regulates Cell Cycle Genes and Potentiates Resistance to Apoptosis. *Sci. Rep.* **8**, 1–12 (2018).

397. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–43 (2013).

398. Kirsten, H. *et al.* Dissecting the genetics of the human transcriptome identifies novel trait-related *trans* -eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum. Mol. Genet.* **24**, 4746–4763 (2015).

399. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

400. J Joo, J. W., Sul, J. H., Han, B., Ye, C. & Eskin, E. Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies. *Genome Biol.* **15**, 1–15 (2014).

401. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

402. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**, 299–309 (2002).

403. Woolfe, A., Mullikin, J. C. & Elnitski, L. Genomic features defining exonic variants that modulate splicing. *Genome Biol.* **11**, R20 (2010).

404. Shahbazian, M. D. & Grunstein, M. Functions of Site-Specific Histone Acetylation and Deacetylation. *Annu. Rev. Biochem.* **76**, 75–100 (2007).

405. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).

406. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).

407. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

408. John, S. *et al.* Genome-Scale Mapping of DNase I Hypersensitivity. in *Current Protocols in Molecular Biology* 21.27.1-21.27.20 (John Wiley & Sons, Inc., 2013). doi:10.1002/0471142727.mb2127s103

409. Altshuler, D. M. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

410. Taliun, D., Gamper, J., Leser, U. & Pattaro, C. Fast Sampling-Based Whole-Genome Haplotype Block Recognition. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **13**, 315–325 (2016).

411. Taliun, D., Gamper, J. & Pattaro, C. Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics* **15**, 10 (2014).

412. Purcell, S. & Chang, C. Plink v1.9.

413. Purcell, S. M. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).

414. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med. Genet.* **7**, 74 (2006).

415. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306**, 636–640 (2004).

416. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human

genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

417. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

418. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–8 (2010).

419. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–37 (2007).

420. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev.* **27**, 1318–1338 (2013).

421. Karmodiya, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, 424 (2012).

422. Lehnertz, B. *et al.* Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA Methylation to Major Satellite Repeats at Pericentric Heterochromatin. *Curr. Biol.* **13**, 1192–1200 (2003).

423. Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**, 3131–41 (2009).

424. Ferrari, K. J. *et al.* Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity. *Mol. Cell* **53**, 49–62 (2014).

425. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

426. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

427. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

428. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

429. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-6 (2014).

430. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805-11 (2015).

431. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92 (2012).

432. Reese, M. G. *et al.* A standard variation file format for human genome sequences. *Genome Biol.* **11**, R88 (2010).

433. MISO Sequence Ontology Browser.

434. Gel, B. *et al.* regioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).

435. Bioconductor Development Team, T. BSgenome.Hsapiens.UCSC.hg38.masked: Full masked genome sequences for Homo sapiens (UCSC version hg38). (2015).

436. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).

437. R_Core_Team. R: A language and environment for statistical computing. (2016).

438. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.* **131**, 217–34 (2012).

439. Fernandez-Rozadilla, C. *et al.* A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics* **14**, 55 (2013).

440. Malapelle, U. *et al.* Less frequently mutated genes in colorectal cancer: evidences from next-generation sequencing of 653 routine cases. *J. Clin. Pathol.* **69**, 767–71 (2016).

441. Williams, C. S. *et al.* ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. *Carcinogenesis* **36**, 710–718 (2015).

442. Hurst, L. D. & Batada, N. N. Depletion of somatic mutations in splicing-associated sequences in cancer genomes. *Genome Biol.* **18**, 1–12 (2017).

443. Cáceres, E. & Hurst, L. D. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* **14**, R143 (2013).

444. Parmley, J. L., Chamary, J. V. & Hurst, L. D. Evidence for Purifying Selection Against Synonymous Mutations in Mammalian Exonic Splicing Enhancers. *Mol. Biol. Evol.* **23**, 301–309 (2006).

445. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science (80-.).* **320**, 1643–1647 (2008).

446. Lee, P. H. *et al.* Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum. Genet.* **137**, 15–30 (2018).

447. Schmid, C. D. & Bucher, P. ChIP-Seq Data Reveal Nucleosome Architecture of Human Promoters. *Cell* **131**, 831–832 (2007).

448. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).

449. Bajpai, R. *et al.* CHD7 cooperates with PBAF to control multipotent neural crest formation. *Nature* **463**, 958–962 (2010).

450. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

451. Watson, J. D. *et al. Molecular biology of the gene*. (Pearson/CSH Press, 2014).

452. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).

453. Carrozza, M. J. *et al.* Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* **123**, 581–592 (2005).

454. Lee, J.-S. & Shilatifard, A. A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutat. Res. Mol. Mech. Mutagen.* **618**, 130–134 (2007).

455. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).

456. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–353 (2006).

457. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

458. Libbrecht, M. W. *et al.* A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *bioRxiv* 086025 (2018). doi:10.1101/086025

459. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

460. Golbabapour, S. *et al.* Gene Silencing and Polycomb Group Proteins: An Overview of their Structure, Mechanisms and Phylogenetics. *OMICS* **17**, 283–296 (2013).

461. Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017).

462. Peng, J. C. & Karpen, G. H. Epigenetic regulation of heterochromatic DNA stability. *Current Opinion in Genetics and Development* **18**, 204–211 (2008).

463. Chen, B., Kenari, N. S. & Libbrecht, M. W. Continuous chromatin state feature annotation of the human epigenome. *bioRxiv* 473017 (2018). doi:10.1101/473017

464. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, (2010).

465. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

466. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

467. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

468. Plesec, T. *et al.* Clinicopathological features of a kindred with SCG5-GREM1–associated hereditary mixed polyposis syndrome. *Hum. Pathol.* **60**, 75–81 (2017).

469. Venkatachalam, R. *et al.* Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int. J. Cancer* **129**, 1635–1642 (2011).

470. Rohlin, A. *et al.* GREM1 and POLE variants in hereditary colorectal cancer syndromes. *Genes, Chromosom. Cancer* **55**, 95–106 (2016).

471. Braks, J. A. M. & Martens, G. J. M. 7B2 is a neuroendocrine chaperone that transiently interacts with prohormone convertase PC2 in the secretory pathway. *Cell* **78**, 263–273 (1994).

472. Yang, H. *et al.* Meta-Analysis of the rs4779584 Polymorphism and Colorectal Cancer Risk. *PLoS One* **9**, e89736 (2014).

473. Yu, J. *et al.* Tumor-derived extracellular mutations of PTPRT /PTPrho are defective in cell adhesion. *Mol. Cancer Res.* **6**, 1106–13 (2008).

474. Wang, Z. *et al.* Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science (80-.).* **304**, 1164–1166 (2004).

475. Laczmanska, I. *et al.* Protein tyrosine phosphatase receptor-like genes are frequently hypermethylated in sporadic colorectal cancer. *J. Hum. Genet.* **58**, 11–15 (2013).

476. Besco, J. A., Frostholm, A., Popesco, M. C., Burghes, A. H. M. & Rotter, A. Genomic organization and alternative splicing of the human and mouse RPTPρ genes. *BMC Genomics* **2**, 1471–1484 (2001).

477. Hynes, N. E. & MacDonald, G. ErbB receptors and signaling pathways in cancer. *Curr. Opin. Cell Biol.* **21**, 177–184 (2009).

478. Shu, X. *et al.* Germline genetic variants in somatically significantly mutated genes in tumors are associated with renal cell carcinoma risk and outcome. *Carcinogenesis* **39**, 752–757 (2018).

479. Walters, J. *et al.* A constitutively active and uninhibitable caspase-3 zymogen efficiently induces apoptosis. *Biochem. J.* **424**, 335–345 (2009).

480. Noble, P. *et al.* High levels of cleaved caspase-3 in colorectal tumour stroma predict good survival. *Br. J. Cancer* **108**, 2097–2105 (2013).

481. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–144 (2013).

482. Domingo, E. *et al.* Somatic POLE proofreading domain mutation, immune response, and prognosis in colorectal cancer: a retrospective, pooled biomarker study. *Lancet Gastroenterol. Hepatol.* **1**, 207–216 (2016).

483. Johnson, R. E., Klassen, R., Prakash, L. & Prakash, S. A Major Role of DNA Polymerase δ in Replication of Both the Leading and Lagging DNA Strands. *Mol. Cell* **59**, 163–175 (2015).

484. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).

485. Fu, S. *et al.* IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics* **34**, 2168–2176 (2018).

486. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).

487. Lee, M. *et al.* Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites. *Am. J. Hum. Genet.* **100**, 751–765 (2017).

488. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).

489. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

490. Frampton, M. J. E. *et al.* Implications of polygenic risk for personalised colorectal cancer screening. *Ann. Oncol.* **27**, 429–434 (2016).

491. Thanki, K. *et al.* Consensus Molecular Subtypes of Colorectal Cancer and their Clinical Implications. *IBBJ* **3**, 105–111 (2017).